JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Volume 19, No. 4
December 2026
ISSN: 1946-1836

In this issue:

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

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

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

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

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Editors

2026 JISARA Editorial Board

Enhancing Healthcare Data Security Through Incentivized Behavioral Cybersecurity

Mary Lind
mary.lind@lsus.edu
Louisiana State University Shreveport
Shreveport, LA 40769

Samuel Tabi
stabi50510@ucumberlands.edu
University of the Cumberlands
Williamsburg, KY 40769

Abstract

Cybercrime has escalated within the healthcare sector, presenting a substantial threat to both operational integrity and patient safety. Notably, 66% of data breaches in healthcare are attributed to providers' failure to identify and address cybersecurity threats. This quantitative correlation study investigated whether factors related to threat avoidance and financial incentives significantly affect healthcare providers' motivation to protect against cyber threats. The study utilized a cross-sectional sample of 107 healthcare practitioners based in the United States. The theoretical framework of the study integrates Carpenter et al.'s refined Technology Threat Avoidance Theory (TTAT) with Jalali et al.'s modified Theory of Planned Behavior (TPB). Employing partial least squares structural equation modeling (PLS-SEM), the study addressed five research questions. The findings indicated that healthcare professionals experienced a heightened sense of control when reliable security technologies were in place. However, variables such as perceived risk, severity, trust in security systems, behavioral control, and financial incentives did not significantly predict motivation for threat avoidance. These results imply that while perceived control is influential, other commonly presumed motivators may not impact cybersecurity behavior as anticipated. Further research should investigate whether factors such as risk propensity, susceptibility, or increased financial incentives can more effectively encourage healthcare providers to adopt robust cybersecurity measures.

Keywords: Cybercrime, healthcare, Technology Threat Avoidance Theory, risk, trust

Recommended Citation: Lind, M.R., Tabi, S., (2026). Enhancing Healthcare Data Security Through Incentivized Behavioral Cybersecurity. *Journal of Information Systems Applied Research and Analytics*, v19(n4) pp 4-14. DOI# https://doi.org/10.62273/UJZB9934

Enhancing Healthcare Data Security Through Incentivized Behavioral Cybersecurity

Mary Lind and Samuel Tabi

1. INTRODUCTION

Data security is of paramount importance in the U.S. healthcare sector because of the protected health information (PHI) contained in electronic health records (EHRs) and patient files (Mbonihankuye et al., 2019; Moore & Frye, 2019; Yeng et al., 2021). These data repositories, which house healthcare records, are particularly attractive targets for cybercriminal intent to commit identity theft (Kaddoura et al., 2021). The motivation for cyberattacks on healthcare organizations stems from the fact that healthcare records encompass critical personal information and sensitive data, which hold greater value on the black market than credit card data (Argaw et al., 2020). Breaches in healthcare data pose significant threats to patient safety and disrupt the normal operation of healthcare institutions (Agrawal et al., 2020; Seh et al., 2020). Such breaches compromise the integrity and confidentiality of patient records, thereby eroding patient trust in healthcare providers (Kaddoura et al., 2021; Yaraghi & Gopal, 2018).

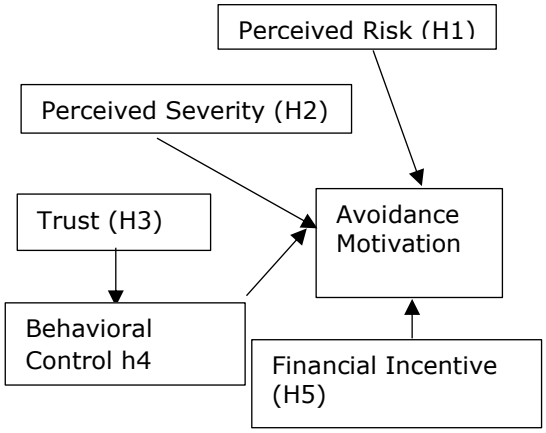
One of the primary causes of data breaches in the healthcare sector is the inability of healthcare providers to detect attacks that target healthcare personnel (Yeng et al., 2021). Information system security professionals within healthcare organizations have employed basic cybersecurity mitigation strategies, such as perimeter fencing, to protect their networks and data (Yeng et al., 2021). Nevertheless, measures such as anti-virus software, intrusion detection and prevention systems, and firewalls have proven ineffective against cyberattacks on healthcare organizations (Yeng et al., 2021). Recent research has indicated that human-related factors, such as inadequate knowledge of healthcare employees or disregard for information security policies, are responsible for the majority of cybersecurity breaches in healthcare organizations (Dong et al., 2021).

Building on the background provided above, this study examines the factors that influence healthcare providers' motivation to protect against cyber threats. The theoretical foundation for this research combines two key frameworks: the Technology Threat Avoidance Theory (TTAT) and an adapted version of the Theory of Planned Behavior (TPB). These theories provide a lens through which to analyze the complex interplay of factors affecting cybersecurity behaviors in healthcare settings. The following section outlines the core components of these theoretical models and their relevance to this study.

2. THEORETICAL BASIS

According to Technology Threat Avoidance Theory (TTAT), when cyber threats are present, users of information systems are inclined to adopt protective measures if they perceive the threat as both probable and potentially severe (Carpenter et al., 2019; Chen & Liang, 2019; Samhan, 2017). If users believe that the threat can be mitigated through specific measures, they are more likely to implement such measures to counteract the threat (Samhan, 2017). In the face of cyber threats, information technology users employ cognitive processes and behaviors to address threats (Chen & Liang, 2019). Users evaluate the severity of the threat and ascertain whether any protective methods are available to counteract it (Chen & Liang, 2019). If a user does not perceive the threat to be sufficiently severe, it may be disregarded (Chen & Liang, 2019). Figure 1 illustrates the TTAT model.

Figure 1 Research Model Based on the TTAT



Trust in Security Technology
The primary element of this study's theoretical framework, as explored in the literature review, is trust in security technology. Trust is an essential aspect of healthcare technology because of the vast datasets managed by organizations (Moore & Frye, 2019, 2020). The digitization of healthcare information systems, including the adoption of Electronic Health Records (EHR) and the Internet of Medical Things (IoMT) in care delivery, presents potential benefits. However, digitization also renders healthcare organizations vulnerable to cyberattacks (Spanakis et al., 2020). Access to and transmission of patient healthcare data via the Internet are insecure because the Internet, as a public network, lacks comprehensive cybersecurity measures (Lee et al., 2021).

Perceived Behavioral Control
Perceived behavioral control is defined as an individual's evaluation of the ease or difficulty involved in executing a particular behavior (Jalali et al., 2020). It functions as a mediator of the effects of attitudes toward the behavior and the influence of subjective norms associated with the behavior (Bosnjak et al., 2020).

Perceived Risk
Perceived risk is conceptualized as the expected likelihood of a negative event occurring (Jalali et al., 2020). In this study, perceived risk refers to the likelihood that cyberattacks may cause harm in healthcare settings. The existing literature suggests that the risks associated with healthcare data are increasing despite significant investments and efforts by healthcare organizations to improve cybersecurity measures (Rachh, 2021). Consequently, the surge in cybercrimes targeting healthcare institutions has resulted in an increase in healthcare data breaches (Argaw et al., 2020).

Financial Incentives
Stakeholders have significant financial incentives in the realms of cybersecurity and healthcare data management. Cybersecurity breaches involving healthcare data result in substantial financial losses for healthcare organizations (Dong et al., 2021; Meisner, 2018). Researchers have observed a notable increase in cyberattacks targeting hospitals and healthcare organizations, which has exacerbated the considerable financial losses experienced by these entities (Gordon et al., 2019).

Perceived Severity
Perceived severity assesses the magnitude of consequences associated with adverse events (Carpenter et al., 2019). The literature review highlights the substantial impact of cybersecurity breaches. When cybercriminals exploit employees' inadequate cybersecurity practices to commit cybercrimes, the repercussions can be severe for patients (Spanakis et al., 2020). In 2018, data breaches in healthcare organizations negatively impacted over six million patients (Semantha et al., 2020). Furthermore, in 2019, more than 41 million patient records were compromised due to healthcare data breaches (Seh et al., 2020).

3. RESEARCH FINDINGS

This study analyzed survey data collected from 107 healthcare providers in the United States, utilizing partial least squares structural equation modeling (PLS-SEM) techniques to examine the data and address five research questions. The theoretical framework for this study is grounded in the technology threat avoidance theory (TTAT) of Carpenter et al. (2019), augmented by prior research conducted by Samhan (2017) and Jalali et al. (2020) concerning information security risk and trust in technology. The 19 survey questions were derived from three distinct surveys conducted by Carpenter et al. (2019), Jalali et al. (2020), and Samhan (2017). The survey incorporated subscales designed to evaluate perceived information security risk, perceived severity, trust in technology reliability, trust in technology functionality, perceived behavioral control, financial incentives, and threat avoidance motivation. Specifically, the instrument included four subscales previously employed by Jalali et al. (2020), which measured perceived information security risk, trust in security technology reliability, trust in security technology functionality, and perceived behavioral control. Additionally, the instrument featured a subscale adapted from Carpenter et al. (2019) to measure perceived severity and a subscale adapted from both Carpenter et al. and Samhan (2017) to assess avoidance motivation. A single item measuring financial motivation was developed, aligning with the focus of this study. The variables were operationalized using 5-point Likert scales. Among the 107 respondents who completed the online survey, 43 were physicians (40.19%), 33 were physician assistants (30.84%), three were nurses (2.80%), and 28 were nurse practitioners (26.17%). Regarding healthcare-related work experience, 5 participants (4.67%) had 0-1 years, 33 participants (30.84%) had 2-5 years, 32 participants (29.91%) had 6-10 years, 20 participants (18.68%) had 11-15 years, and 17 participants (15.89%) had more than 15 years.

All respondents were aged between 25 and 75 years. Specifically, 42 participants (39.25%) were aged 25 to 34, 47 (43.93%) were aged 35 to 44, 11 (10.28%) were aged 45 to 54, 5 (4.67%) were aged 55 to 64, and 2 (1.87%) were aged 65 to 74. The sample was relatively balanced in terms of sex, with 59 males (55.14%) and 48 females (44.86%). Appendix A contains the survey items used.

The reliability evaluation presented in Table 1 indicates that all constructs demonstrated reliability.

Table 1 Construct Reliability

Construct	Cronbach's α	Composite Reliability
Perceived information security risk	0.925	0.938
Perceived severity	0.941	0.965
Trust in security technology - reliability	0.881	0.926
Trust in security technology - functionality	0.905	0.940
Perceived behavioral control	0.887	0.930
Avoidance motivation	0.880	0.941

Note. All demonstrated high reliability

Variance Inflation Factor (VIF) analysis revealed that the constructs were not multicollinear. Furthermore, by employing the Fornell-Larcker criterion, the constructs exhibited discriminant validity.

The hypotheses were assessed using PLS-SEM, with the findings detailed in Table 2 and Figure 2 in Appendix B. Additionally, Table 3 presents the outcomes of the hypotheses derived from the PLS-SEM analysis.

Table 2 Results of the Path Analysis

Path	β	t	p
PIS -> AM	-0.075	0.331	0.741
PS -> AM	-0.028	0.151	0.880
TTR -> PBC	0.488	3.239	0.001*
TTF -> PBC	0.238	1.603	0.109
PBC -> AM	0.135	1.404	0.160
Fin -> AM	0.102	0.952	0.341

Table 3 Hypothesis Results Summary

H	Variable Relationship	Result
1	Perceived information security risk -> Avoidance motivation	Rejected
2	Perceived severity -> Avoidance motivation	Rejected
3A	Trust in security technology-reliability -> Perceived behavioral control	Supported
3B	Trust in security technology-functionality -> Perceived behavioral control	Rejected
4	Perceived behavioral control -> Avoidance motivation	Rejected
5	Financial incentive -> Avoidance motivation	Rejected

The analysis of the five research questions yielded mixed results regarding the factors influencing healthcare providers' motivation to protect against cyber threats. While perceived behavioral control was significantly predicted by trust in security technology reliability, the data did not support other hypothesized relationships. Specifically, as expected, perceived risk, perceived severity, trust in security technology functionality, and financial incentives did not significantly predict threat avoidance motivation. These findings suggest that the factors influencing cybersecurity behaviors among health care providers may be more complex than initially theorized. The following section explores the implications of these results, discusses potential explanations for the unexpected findings, and proposes directions for future research to examine the drivers of cybersecurity practices in healthcare settings.

4. CONCLUSIONS and IMPLICATIONS

The lack of a significant correlation between perceived risk and healthcare providers' motivation to avoid cyber threats may be attributed to factors such as participants' workload or workplace culture (Seh et al., 2020). Excessive workloads can adversely impact the cognitive capacities of healthcare employees, thereby reducing their ability to implement measures to safeguard healthcare data (Seh et al., 2020). Furthermore, overburdened healthcare employees may become dissatisfied and choose not to adhere to their organizations' security policies (ISSPs; Jalali et al., 2020). Another potential explanation for the absence of a significant relationship between perceived risk and avoidance motivation is a lack of security awareness.

Scholarly literature has linked security awareness to the support of top management (Dong et al., 2021). Consequently, a workplace culture that fails to prioritize cybersecurity may result in employees lacking sufficient understanding of the importance of protecting against cybersecurity threats.

Although perceived severity did not significantly predict healthcare providers' avoidance motivation to protect against cyber threats in the present study, previous research has demonstrated significant findings. For example, Carpenter et al. (2019) identified a statistically significant correlation between employees' perception of the severity of a cyber threat and their avoidance motivation. One possible explanation for the absence of a significant result for Research Question Two is low cybersecurity awareness. Grassegger and Nedbal (2021) emphasized the importance of cybersecurity awareness as a precursor to compliance. If an employee does not perceive a threat as severe, they may disregard it, particularly if compliance is perceived as burdensome or inconvenient. Rostami et al. (2020) observed that Information Security Policies (ISSPs) can create stress and burdens on employees. Consequently, threats may be disregarded as insignificant if they allow employees to avoid onerous cybersecurity measures.

The initial set of hypotheses examined trust as predicated on reliability. The data analysis revealed that trust in technology, grounded in reliability, significantly predicted healthcare providers' perceived behavioral control in safeguarding against cyber threats. Conversely, the association between trust in security technology based on functionality and healthcare providers' perceived behavioral control in mitigating cyber threats was found to be insignificant. Therefore, a healthcare provider's perceived behavioral control in defending against cyber threats is primarily influenced by their trust in security technology's consistent success in providing protection. This finding corroborates with the results reported by Jalali et al. (2020).

Perceived behavioral control pertains to participants' beliefs regarding their ability to influence cybersecurity outcomes through their compliance with Information Systems Security Policies (ISSP). The results pertaining to Research Question Four of this study indicate that perceived behavioral control did not exhibit a significant relationship with healthcare providers' motivation to avoid cyber threats, leading to the retention of the null hypothesis. This outcome does not align with the theoretical framework of this study or previous research findings (Bosnjak et al., 2020; Jalali et al., 2020).

A financial incentive is defined as a monetary bonus offered to healthcare providers with the aim of positively influencing their motivation to avoid cyber threats. However, the statistical analysis did not reveal a significant relationship between financial incentives and the avoidance motivations of healthcare providers. Consequently, financial incentives do not significantly predict healthcare providers' motivation to protect against cyber threats. The literature indicates that the financial costs associated with healthcare data breaches are substantial (Argaw et al., 2020; Dong et al., 2021; Seh et al., 2020). Cyber breaches of healthcare data result in considerable financial losses for healthcare organizations (Dong et al., 2021; Meisner, 2018). These financial costs may include ransoms paid to recover data, fines for HIPAA violations, or revenue losses due to reputational damage (Allen, 2021; Argaw et al., 2020; M. C. Williams et al., 2020). Such financial implications suggest that offering incentives to employees to enhance compliance could be a potentially beneficial strategy.

5. FUTURE RESEARCH

Further research should investigate the differences in risk propensities between healthcare providers and other occupational groups. The healthcare sector is characterized by high levels of stress and demands, making it particularly suitable for individuals with a propensity for risk-taking. Understanding the relationship between risk-taking behavior and cyber threat avoidance could elucidate the factors that influence the decision to implement protective measures for information assets (Carpenter et al., 2019). A comparative research design would enable researchers to assess the risk propensity of various healthcare providers and determine whether specific job roles attract individuals who are more inclined to neglect information security. Furthermore, future research could explore the differences in information security attitudes between the healthcare sector and other industries such as manufacturing, education, retail, and finance. According to TTAT, when a cyber threat is present, the motivation of an information system user to employ a safeguard measure against this threat is contingent upon the user's perception

of the threat (Carpenter et al., 2019; Chen & Liang, 2019; Samhan, 2017). The theoretical framework of this study did not significantly predict the motivations and behaviors of healthcare providers regarding cyber threat avoidance. Other factors are likely to influence threat avoidance motivations among healthcare providers. Additionally, workplace overload and insufficient security training may have affected participants' avoidance motivations (Dong et al., 2021; Seh et al., 2020). Employing an alternative theoretical framework that incorporates additional variables such as fear appeals or risk susceptibility would provide further insight into the antecedents of ISSP compliance motivations and behaviors.

Based on this study's findings, financial incentives do not significantly predict a healthcare provider's avoidance motivation to protect against cyber threats. If an incentive for a physician is very small compared to the physician's salary, the incentive will not significantly influence the physician's behavior (Vilendrer et al., 2021). Vilendrer et al. (2021) noted that for an employee to experience enhanced motivation as a result of a financial incentive, the incentive must be substantial, approximately 10% to 20% of the healthcare professional's salary. Additional research should be conducted to determine whether increasing the incentive threshold for healthcare providers would significantly change their threat avoidance motivation.

Yoo et al. (2018) noted that individuals who exhibit psychological ownership of computing devices and the Internet in their homes initiate and display proactive cybersafety behaviors. If an employee of an organization experiences psychological ownership of the organizational resources where they work, they will feel compelled to take measures to safeguard those resources (Verkijika, 2020). Further research should examine health care professionals' psychological ownership of their organizations' health information systems. Healthcare professionals who exhibit psychological ownership of health information systems in their organizations might be more motivated by financial incentives to avoid cyber threats (Verkijika, 2020).

6. REFERENCES

Agrawal, A., Pandey, A., Baz, A., Alhakami, H., Alhakami, W., Kumar, R., & Khan, A. (2020). Evaluating the security impact of healthcare web applications through fuzzy based hybrid approach of multi-criteria decision-making analysis. *IEEE Access, 8*, 135770–135783. https://doi.org/10.1109/ACCESS.2020.3010729

Allen, A. (2021). HIPAA at 25—A work in progress. *The New England Journal of Medicine, 384*(23), 2169–2171. https://doi.org/10.1056/NEJMp2100900

Argaw, S. T., Troncoso-Pastoriza, J. R., Lacey, D., Florin, M.-V., Calcavecchia, F., Anderson, D., Burleson, W., Vogel, J.-M., O'Leary, C., Eshaya-Chauvin, B., & Flahault, A. (2020). Cybersecurity of hospitals: Discussing the challenges and working towards mitigating the risks. *BMC Medical Informatics and Decision Making, 20*, Article 146. https://doi.org/10.1186/s12911-020-01161-7

Bosnjak, M., Ajzen, I., & Schmidt, P. (2020). The theory of planned behavior: Selected recent advances and applications. *Europe's Journal of Psychology, 16*(3), 352–356. https://doi.org/10.5964/ejop.v16i3.3107

Carpenter, D., Young, D. K., Barrett, P., & McLeod, A. J. (2019). Refining technology threat avoidance theory. *Communications of the Association for Information Systems, 44*, 380–407. https://doi.org/10.17705/1CAIS.04422

Chen, D., & Liang, H. (2019). Wishful thinking and IT threat avoidance: An extension to the technology threat avoidance theory. *IEEE Transactions on Engineering Management, 66*(4), 552–567. https://doi.org/10.1109/TEM.2018.2835461

Dong, K., Ali, F., Dominic, D., & Ali, A. (2021). The effect of organizational information security climate on information security policy compliance: The mediating effect of social bonding towards healthcare nurses. *Sustainability, 13*(5), Article 2800. https://doi.org/10.3390/su13052800

Gordon, J., Wright, A., Aiyagari, R., Corbo, L., Glynn, R. J., Kadakia, J., Kufahl, J., Mazzone, C., Noga, J., Parkulo, M., Sanford, B., Scheib, P., & Landman, A. B. (2019). Assessment of employee susceptibility to phishing attacks at US health care

institutions. *JAMA Network Open,* 2(3), Article e190393. https://doi.org/10.1001/jamanetworkopen.2019.0393

Grassegger, T., & Nedbal, D. (2021). The role of employees' information security awareness on the intention to resist social engineering. *Procedia Computer Science, 181*, 59–66. https://doi.org/10.1016/j.procs.2021.01.103

Jalali, M., Bruckes, M., Westmattelmann, D., & Schewe, G. (2020). Why employees (still) click on phishing links: Investigation in hospitals. *Journal of Medical Internet Research, 22*(1), Article e16775. https://doi.org/10.2196/16775

Kaddoura, S., Haraty, R., Al Kontar, K., & Alfandi, O. (2021). A parallelized database damage assessment approach after cyberattack for healthcare systems. *Future Internet, 13*(4), Article 90. https://doi.org/10.3390/fi13040090

Lee, T.-F., Chang, I.-P., & Kung, T.-S. (2021). Blockchain-based healthcare information preservation using extended chaotic maps for HIPAA privacy/security regulations. *Applied Sciences, 11*(22), Article 10576

Liang, H., & Xue, Y. (2009). Avoidance of information technology threats: A theoretical perspective. *MIS Quarterly, 33*(1), 71–90. https://doi.org/10.2307/20650279

Mbonihankuye, S., Nkunzimana, A., & Ndagijimana, A. (2019). Healthcare data security technology: HIPAA compliance. *Wireless Communications and Mobile Computing, 2019*, Article 192749

Meisner, M. (2018). Financial consequences of cyber-attacks leading to data breaches in the healthcare sector. *Copernican Journal of Finance & Accounting, 6*(3), 63–73. https://doi.org/10.12775/CJFA.2017.017

Moore, W., & Frye, S. (2019). Review of HIPAA, part 1: History, protected health information, and privacy and security rules. *Journal of Nuclear Medicine Technology*, *47*(4), 269–272. https://doi.org/10.2967/jnmt.119.227819

Moore, W., & Frye, S. (2020). Review of HIPAA, part 2: Limitations, rights, violations, and role for the imaging technologist. *Journal of Nuclear Medicine Technology, 48*(1), 17–23. https://doi.org/10.2967/jnmt.119.227827

Rachh, A. (2021). A study of future opportunities and challenges in digital healthcare sector: Cyber security vs. crimes in digital healthcare sector. *Asia Pacific Journal of Health Management*, *16*(3), 7–15. https://doi.org/10.24083/apjhm.v16i3.957

Rostami, E., Karlsson, F., & Kolkowska, E. (2020). The hunt for computerized support in information security policy management: A literature review. *Information Management & Computer Security, 28*(2), 215–259. https://doi.org/10.1108/ICS-07-2019-0079

Samhan, B. (2017). Security behaviors of healthcare providers using HIT outside of work: A technology threat avoidance perspective. In *2017 8th International Conference on Information and Communication Systems* (pp. 342–347). IEEE. https://doi.org/10.1109/IACS.2017.7921995

Seh, A. H., Zarour, M., Alenezi, M., Sarkar, A. K., Agrawal, A., Kumar, R., & Khan, R. A. (2020). Healthcare data breaches: Insights and implications. *Healthcare*, *8*(2), Article 133. https://doi.org/10.3390/healthcare8020133

Semantha, F. H., Azam, S., Yeo, K. C., & Shanmugam, B. (2020). A systematic literature review on privacy by design in the healthcare sector. *Electronics, 9*(3), Article 452. http://dx.doi.org/10.3390/electronics9030452

Si, H., Shi, J.-G., Tang, D., Wen, S., Miao, W., & Duan, K. (2019). Application of the theory of planned behavior in environmental science: A comprehensive bibliometric analysis. *International Journal of Environmental Research and Public Health, 16*(15), Article 2788. https://doi.org/10.3390/ijerph16152788

Spanakis, E. G., Bonomi, S., Sfakianakis, S., Santucci, G., Lenti, S., Sorella, M., Tanasache, F. D., Palleschi, A., Ciccotelli, C., Sakkalis, V., & Magalini, S. (2020). Cyber-attacks and threats for healthcare: A multi-layer thread analysis. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society* (pp. 5705–5708). IEEE. https://doi.org/10.1109/EMBC44109.2020.9

Verkijika, S. (2020). Employees' cybersecurity behaviour in the mobile context: The role of self-efficacy and psychological ownership. In *2020 2nd International Multidisciplinary Information Technology and Engineering Conference* (pp. 1–5). IEEE. https://doi.org/10.1109/IMITEC50163.2020.9334097

Vilendrer, S., Brown-Johnson, C., Kling, S. M. R., Veruttipong, D., Amano, A., Bohman, B., Daines, W. P., Overton, D., Srivastava, R., & Asch, S. M. (2021). Financial incentives for medical assistants: A mixed-methods exploration of bonus structures, motivation, and population health quality measures. *Annals of Family Medicine, 19*(5), 427–436. https://doi.org/10.1370/afm.2719

Williams, C. M., Chaturvedi, R., & Chakravarthy, K. (2020). Cybersecurity risks in a pandemic. *Journal of Medical Internet Research, 22*(9), Article e23692. https://doi.org/10.2196/23692

Yaraghi, N., & Gopal, R. (2018). The role of HIPAA omnibus rules in reducing the frequency of medical data breaches: Insights from an empirical study. *The Milbank Quarterly, 96*(1), 144–166. https://doi.org/10.1111/1468-0009.12314

Yeng, P. K., Szekeres, A., Yang, B., & Snekkenes, E. A. (2021). Mapping the psychosocial, cultural aspects of healthcare professionals' information security practices: Systematic mapping study. *JMIR Human Factors, 8*(2), Article e17604. https://doi.org/10.2196/17604

Yoo, C. W., Sanders, G. L., & Cerveny, R. P. (2018). Exploring the influence of flow and psychological ownership on security education, training and awareness effectiveness and security compliance. *Decision Support Systems*, *108*, 107–118. https://doi.org/10.1016/j.dss.2018.02.009

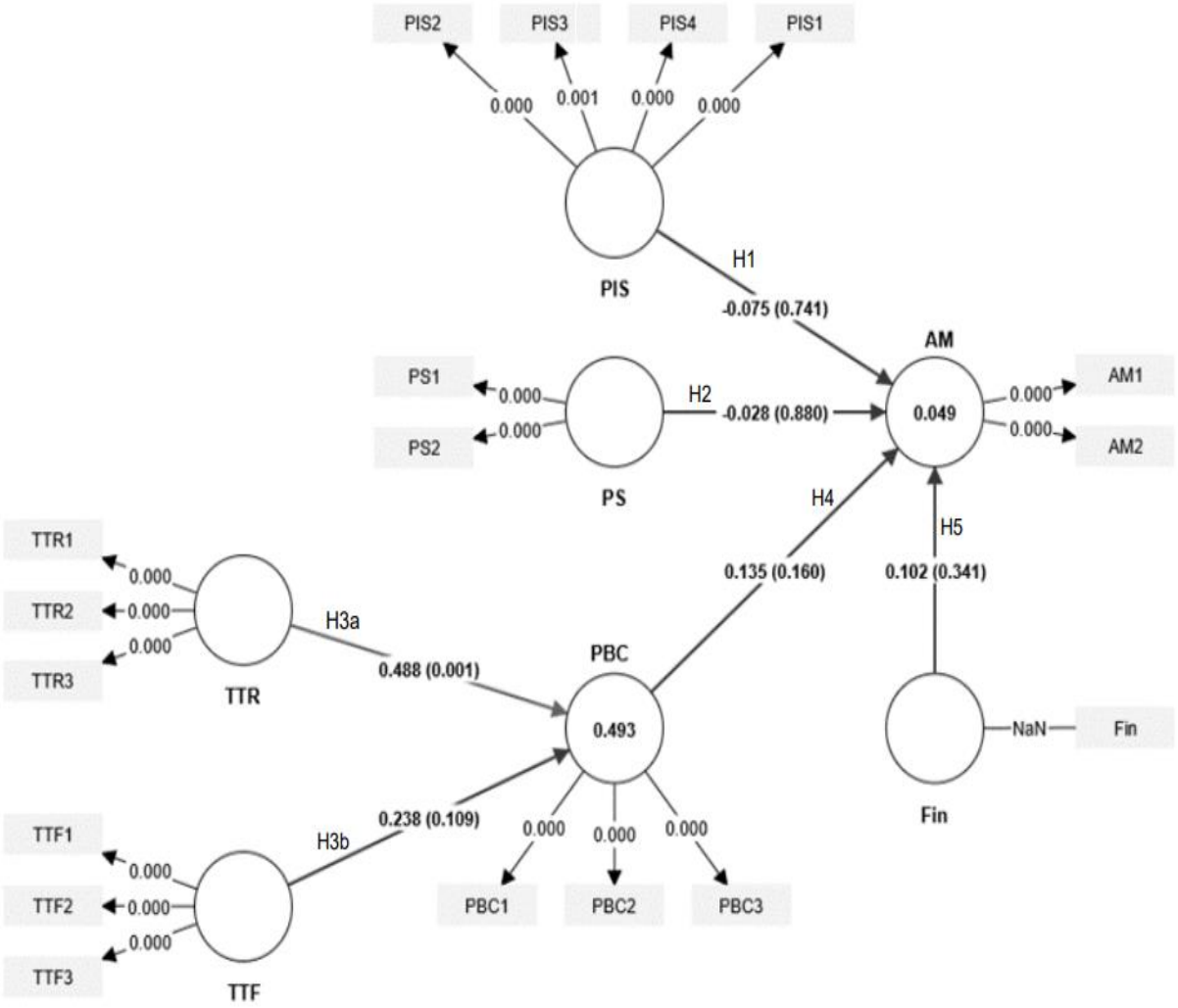
Appendix A

Data Collection Instrument

Item	Question	Construct	Scale	Source
1	At my workplace, the risk to my computer and data from Internet security breaches is:	Perceived information security risk	A	Jalali et al. (2020)
2	At my workplace, the likelihood that my computer will be disrupted due to Internet security breaches within the next 12 months is:	Perceived information security risk	A	Jalali et al. (2020)
3	At my workplace, the chance that my computer will fall a victim to an Internet security breach is:	Perceived information security risk	A	Jalali et al. (2020)
4	At my workplace, the vulnerability of my computer and data to Internet security risks is:	Perceived information security risk	A	Jalali et al. (2020)
5	My personal information collected by malware at my place of work could be used to commit crimes against me.	Perceived severity	A	Carpenter et al. (2019)
6	Health records of patients collected by malware at my place of work could be used to commit crimes against patients.	Perceived severity	A	Carpenter et al. (2019)
7	The cybersecurity software at my workplace (e.g., antivirus and firewall) is reliable.	Trust in technology –reliability	B	Jalali et al. (2020)
8	The cybersecurity software at my workplace does not fail me.	Trust in technology –reliability	B	Jalali et al. (2020)
9	The cybersecurity software at my workplace provides accurate service.	Trust in technology –reliability	B	Jalali et al. (2020)
10	The cybersecurity software at my workplace has the functionality I need.	Trust in technology –functionality	B	Jalali et al. (2020)
11	The cybersecurity software at my workplace has the features required for my tasks.	Trust in technology – functionality	B	Jalali et al. (2020)
12	The cybersecurity software at my workplace has the ability to do what I want it to do.	Trust in technology – functionality	B	Jalali et al. (2020)
13	I am able to follow the cybersecurity policies and procedures and technologies (e.g., antivirus, or other products).	Perceived behavior control	B	Jalali et al. (2020)
14	I have the resources and knowledge to follow the policies and procedures and use the cybersecurity technologies.	Perceived behavior control	B	Jalali et al. (2020)

15	I have adequate training to follow the policies and procedures and use cybersecurity technologies.	Perceived behavior control	B	Jalali et al. (2020)
16	I intend to use anti-spyware software to avoid spyware.	Threat avoidance motivation	B	Carpenter et al. (2019), Samhan, (2017)
17	I predict I would use anti-spyware software to avoid spyware.	Threat avoidance motivation	B	Carpenter et al. (2019), Samhan, (2017)
18	I plan to use anti-spyware software to avoid spyware.	Threat avoidance motivation	B	Carpenter et al. (2019), Samhan, (2017)
19	I would follow the cybersecurity policies and procedures and technologies (e.g., antivirus, or other products) more if there were a financial incentive of about 10% to 20% of my salary.	Financial motivation	B	Developed based on this study's focus

Appendix B PLS-SEM Model



Designing A Recommender System with Hybrid Models

Dmytro Dobrynin
dd1503@uncw.edu

Yao Shi
shiy@uncw.edu

Jeff Cummings
cummingsj@uncw.edu

Gulustan Dogan
dogang@uncw.edu

University of North Carolina Wilmington
Wilmington,NC 28405

Abstract

The rapid growth of digital media content makes it difficult to provide timely, relevant, and personalized recommendations. In isolation, traditional recommender systems are effective, but they often struggle with data sparsity, the cold-start problem, diversity, and adaptability to changing user preferences. Systems that accurately interpret user behaviour and item characteristics are essential. Addressing these challenges requires strategies that go beyond conventional collaborative or content-based filtering alone. This study aims to address the following research question: How can hybrid approaches integrating traditional recommendation techniques with modern machine learning methods improve the personalization, diversity, and resilience of a recommender system? To this end, we developed hybrid movie recommendation models by combining collaborative filtering with content-based analysis using NLP and two-tower neural network architectures. The collaborative filtering components utilize matrix factorization to uncover latent user preferences, while natural language processing techniques extract semantic features from movie descriptions to enhance content understanding. Neural Retrieval-Ranking models help to further refine recommendations by learning compact representations of users and items, enabling efficient and adaptive candidate selection. The evaluation methodology included both offline algorithmic performance measurement and user-centered assessments. The findings demonstrate the efficacy of selected hybrid strategies for personalized recommendations across similar application domains.

Keywords: Recommender Systems, Hybrid Filtering, Content-based Filtering, Collaborative Filtering, Natural Language Processing.

Recommended Citation: Dobrynin, D., Shi, Y., Cummings, J., Dogan, G., (2026). Designing A Recommender System with Hybrid Models. *Journal of Information Systems Applied Research and Analytics*, v19(n4) pp 15-25. DOI# https://doi.org/10.62273/ZUUE1587

A Design of Recommender Systems with Hybrid Models.

Dmytro Dobrynin, Yao Shi, Jeff Cummings and Gulustan Dogan

1. INTRODUCTION

With the widespread adoption of the Internet in homes and on mobile devices, the issue of information and media content oversaturation has become increasingly prominent. The sheer volume of available choices now far exceeds users' needs, making it challenging to filter and prioritize content for efficient and timely delivery. Recommender systems address those challenges by leveraging user-specific data to identify and deliver the most relevant content—particularly in streaming services and online retail platforms—ensuring users receive what they truly need.

Many commercial enterprises, such as Amazon, TripAdvisor, and IMDb, have successfully integrated recommender systems into their platforms. Unlike Netflix, which primarily focuses on suggesting films and TV series, these companies offer a broader range of products and services, supported by more diverse catalogues. This highlights the adaptability and versatility of recommender systems, which are not limited to a single domain. They can effectively guide users toward discovering books, exploring travel destinations, or even adopting innovative new technologies, all tailored to individual preferences.

However, designing and evaluating recommender systems remain persistent challenges. Several researchers (Cremonesi et al., 2010; Konstan & Riedl, 2012) argue that users are less concerned with precise rating predictions and more interested in whether the system can effectively recommend items that align with their needs and preferences. Moreover, some deep learning models, such as NeuMF and DeepFM, while powerful, are often overly complex and require large datasets to perform optimally. These models may not outperform lightweight alternatives in resource-constrained environments, where low integration and deployment time are critical. In addition, large language models (LLMs), as emerging AI techniques, demonstrate advanced capabilities in understanding linguistic complexities. Yet, they cannot fully replace the existing recommender systems, as users' preferences are derived not only from textual data but also from behavioral patterns (Li et al., 2023; Zhao et al., 2024).

This disconnect highlights the need for design and evaluation metrics that better reflect real-world user satisfaction and engagement within resource-constrained environments.

Building on the foundational concepts, definitions, and applications of recommender systems, this research aims to explore and evaluate methodologies for designing, developing, and assessing recommender systems using hybrid filtering techniques. The study places particular emphasis on implementing and comparing the accuracy and effectiveness of several custom models, guided by the principle of combining well-discovered and analysed approaches to achieve peak performance (Burke, 2002). Hybrid models, which synthesize multiple recommendation strategies, usually show an ability to mitigate individual model limitations and make the resulting system more robust and, in the case of recommenders, provide more accurate, relevant, effective, and manipulation-resistant systems to satisfy user needs (Konstan & Riedl, 2012). This research addresses common challenges in recommender systems, including the cold-start problem, new-user limitations, and the need for frequent model retraining, to enhance overall system reliability and user satisfaction.

This research evaluates the performance of the mentioned models and concludes which model turned out to be the most accurate and able to satisfy more users than others. Additionally, it identifies potential improvements, as well as the enhancements that the proposed approach could offer to existing and offered hybrid models in the future.

The study is organised into the following sections. First, the literature review section introduces the popular methods of building recommender systems, their disadvantages and advantages, and how hybrid approaches can be applied to address emerging problems. The following sections, prototype design, implementation, and evaluation, demonstrate the conception of each prototype, the implementation process for the prototypes, and

the procedure for testing and assessing the prototypes. Finally, the evaluation findings, potential improvements and overall outcome are presented in the discussion, future work, and conclusion sections.

2. LITERATURE REVIEW

There are three main traditional approaches to building recommender systems: collaborative filtering, content-based filtering, and hybrid filtering. Each approach is distinguished by its underlying logic, data used and methodology for generating recommendations.

Collaborative Filtering

Collaborative filtering (CF) is a widely used recommendation technique that suggests items to users by leveraging the preferences and behaviours of other users with similar tastes (Isinkaye et al., 2015; Papadakis et al., 2022).

CF is typically implemented using two primary approaches: user-based and item-based methods. User-based CF identifies similar users based on their historical preferences and recommends items that those users have liked (Isinkaye et al., 2015). In contrast, item-based CF calculates similarities between items based on user ratings and recommends items that are similar to those the target user has rated highly. Both approaches rely on similarity measures, such as cosine similarity or Pearson correlation, to determine the level of alignment between users or items.

CF systems offer several advantages. One of the key strengths is the ability to generate serendipitous recommendations. This enhances user engagement and is particularly effective in domains where item content is difficult to analyze. Additionally, CF systems improve over time as user interactions accumulate, enriching the dataset and enhancing recommendation accuracy. However, CF also faces significant challenges. A prominent issue is a cold-start problem, which occurs when there is insufficient data about new users or new items, resulting in many missing values in the User-Item matrix. Data sparsity is a closely related problem, where users rate only a small number of items.

Content-Based Filtering

Content-based filtering (CBF) is a recommendation technique that provides personalised item suggestions by analysing the intrinsic attributes of items and aligning them with a user's historical preferences (Javed et al., 2021; Thannimalai & Zhang, 2021).

CBF operates by comparing new items to the items that a user has previously rated positively. This process typically involves the application of various algorithmic models to assess the similarity between items. Vector Space Models (VSM), such as Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA), are commonly used to represent textual content and project the two documents onto one of the models to calculate their similarities (Falk, 2019). These methods enable the system to predict which items are likely to appeal to a user based on their prior interactions.

CBF offers key advantages, including user independence and transparency. Unlike CF, CBF relies exclusively on the user's preferences, enabling personalized recommendations based on individual interaction history. It also addresses the cold-start problem for new items, a common limitation in CF. However, CBF has notable limitations. One major limitation is overspecialization, where the system tends to recommend items or similar ones that the user has interacted with, thereby reducing diversity and novelty. Moreover, CBF, similar to CF, struggles with data sparsity, limiting its effectiveness for users with minimal interaction history.

Hybrid Filtering Model

Given the respective strengths and limitations of the above recommender systems construction techniques, it is both practical and beneficial to develop a system that would combine them to achieve better performance with fewer drawbacks of any individual one (Burke, 2002; Thorat et al., 2015). These systems are known as hybrid systems. To address the limitations of individual strategies and leverage their strengths, hybrid systems combine content-based filtering and collaborative filtering.

Burke (2002) classifies hybrid recommender systems into seven main types: weighted, switching, mixed, feature combination, feature augmentation, cascade, and meta-level. Each type represents a distinct strategy for integrating multiple recommendation techniques. Among these, the cascade model is particularly notable for its efficiency and precision. In cascade hybridization, one recommendation method is used to generate an initial, coarse list of candidate items, which is then refined by a second method. This sequential approach avoids applying complex or computationally intensive techniques to items that are either clearly irrelevant or already well-

differentiated. As a result, cascade hybrids enhance both performance and accuracy by focusing on refinement efforts only where they are most needed.

While hybridization helps alleviate cold-start limitations—particularly through the two-tower component's ability to use auxiliary features—the collaborative filtering component still requires retraining to fully incorporate new users or items.

Two-Tower Model

The two-tower model separates user and item features into two independent networks. This structure allows each network to specialize, enabling the model to learn more granular, feature-specific representations for users and items (Wang et al., 2025; Wortz & Totten, 2023). The two-tower architecture represents a significant advancement in recommender systems, addressing the limitations of traditional approaches that often rely solely on CF or CBF (Yang et al., 2020). Unlike single-model systems that focus on either user preferences or item similarity, this architecture employs two neural network "towers"—one for user features and another for item features (Varsha, 2024). This design offers several advantages. Its decoupled embedding structure enhances scalability, allowing independent and efficient training of user and item towers. It also supports diverse data integration, leveraging both structured and unstructured information to improve generalization. Additionally, the architecture enables online learning by precomputing embeddings, allowing rapid updates for new users or items without retraining the entire model. This dynamic adaptability makes the two-tower framework both efficient and responsive (Lee & Cho, 2023).

3. PROTOTYPE DESIGN

This study attempts to integrate all the above techniques into a multistage hybrid recommender system based on publicly available movie datasets. It leverages the strengths of both CF and CBF, while also incorporating the novel opportunities presented by the two-tower neural network for retrieval and ranking tasks. Our hybridisation strategies are demonstrated through the three prototypes as exhibited in Figure 1:

Prototype 1: "Simple Retrieval, Ranking, SVD and LDA"

The initial approach involves constructing a Retrieval-Ranking system based on a two-tower architecture, utilizing a limited set of features such as user and movie identifiers, along with user-movie ratings. This system is further enhanced through the integration of Singular Value Decomposition (SVD) (Klema & Laub, 1980), serving as the CF submodel, and Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Chang et al., 2023; Jelodar et al., 2019), serving as the CBF submodel.

During the retrieval phase, an initial set of candidates will be selected from the entire pool of available items, and any items that the user is not interested in will be weeded out. This step can help to reduce the number of candidates — from potentially millions or tens of thousands of items to a more manageable subset. The following Ranking submodel not only reduces the number of candidate items further but also provides an estimated rating for each item, enabling them to be sorted accordingly. SVD allows to expand outputs of the Ranking submodel with items similar to those already highly rated and recommended to watch. By computing the vectors' element-wise sum (dot product) of the user embedding and the item embedding, the model generates a score for each potential item for a specific user, allowing it to generate the top k items with the highest score for the chosen user. Embedding sizes and learning rates for each model are selected accordingly to reduce training time, omit overfitting, and make the model accuracy on the test dataset better. Search space for embedding size was between 16 and 128 with step of 8. Search space for learning rate was between 1e-4 to 1e-1 with "log" sampling. For tuning, Random Search from keras_tuner was used.

After, LDA will function as a content-based model that analyses item descriptions or further features in future. Before implementing the LDA submodel, it is necessary to select a proper number of topics K. The Value of K is selected by trying a range of topics from 2 to 100 and measuring coherence. After K was selected and the movie's textual descriptions were divided into topics, it will be possible to visualise those topics and the frequency of words in them.

Prototype 2: "SVD and LDA"

The second hybrid approach involves a simpler sequential flow where SVD is only followed by LDA. Initially, SVD will be applied to extract latent user preferences swiftly and find items similar to the ones with the highest ranking. Subsequently, LDA will refine these suggestions by analysing the semantic content of the movie

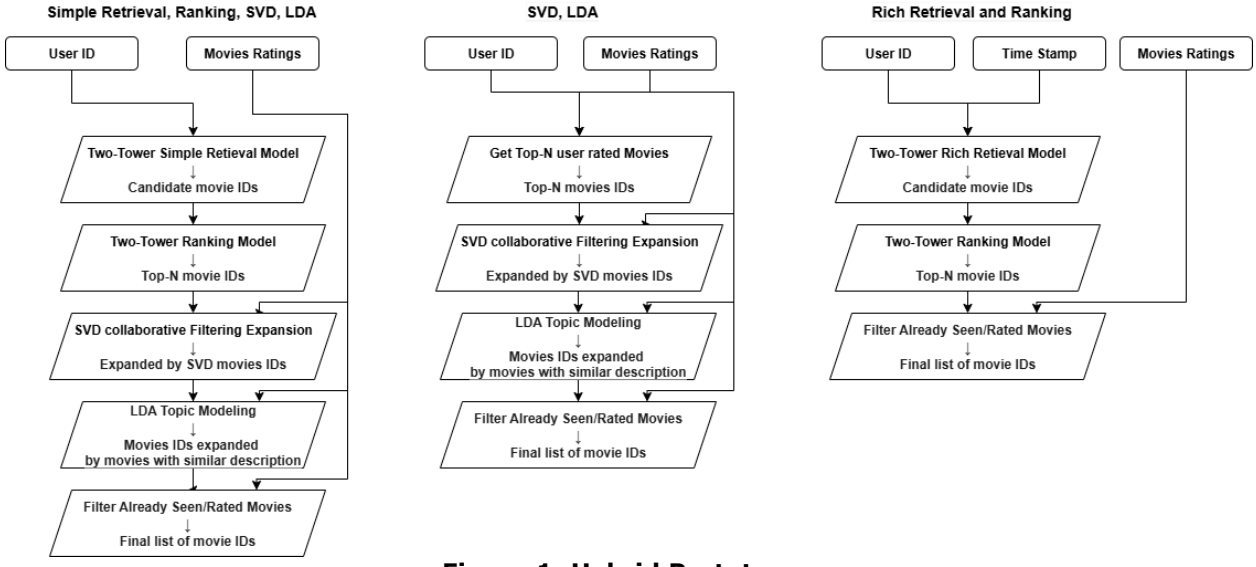


Figure 1: Hybrid Prototypes

descriptions. For example, if a user has shown an affinity for certain genres or themes through their interaction history, LDA can identify and recommend items that include similar topics

This two-tiered method will allow us to address the shortcomings of each submodel. While SVD may overlook nuanced content features, LDA may compensate for them by adding a layer of contextual understanding, resulting in a more robust recommendation engine.

Prototype 3: "Rich Retrieval and Ranking"
The last approach enhances the first strategy by incorporating additional features into the first prototype model. These features include the timestamp and the movie's title. The architecture of the Retrieval submodel with more features. In this step, the process remains iterative after generating potential movie candidates through the Retrieval submodel with more features. Followed by ranking them using the Ranking submodel, the recommendations will be refined. This will not only allow for more personalised recommendations but also cater to the diverse range of user profiles that are prevalent in business settings.

After each stage in the studying techniques, it is essential to ensure that the films in the resultant recommendation list have not been previously viewed or reviewed by the selected user for whom the to-watch list is generated. This approach allows for evaluating how effectively the final prototypes suggest similar items without re-recommending content. When training two-tower submodels, it is crucial to prevent them from merely memorizing data from the training set; otherwise, the models may only output movies that have already been viewed or rated, leading to repetitive recommendations of items from the user's existing watch list.

4. PROTOTYPE IMPLEMENTATION

The project implementation began with collecting and processing data that was used for recommender models and their training. Most of the work was based on data found on the MovieLens website, which contains around 100 thousand ratings for 9 thousand movies (Harper & Konstan, 2015). Additional data from the same site was used to supplement the dataset for the LDA submodel. Only a table with information about 60 thousand movies from MovieLens 25M dataset was used.

Textual descriptions or plot summaries were extracted from every movie available on the IMDb website using a developed web scraper. To make the textual descriptions ready to be used for the described LDA submodel it was needed to clean texts from special symbols and punctuation and proceed to removing stop words, stemming and tokenizing the processed corpus of texts.

To find the proper K value, a series of experiments was executed to find the relationship between topic coherence and target value. The results of those experiments are

depicted in Figure 2. Based on the retrieved dependency between coherence and the number of topics, it was decided to select a value K equal to 20. The reason K = 20 was chosen is that we ran experiments varying the number of LDA topics from 2 to 100, measured the topic coherence for each, and selected the value that produced the best trade-off between interpretability and coherence. According to the text, the coherence curve in Figure 2 shows that K = 20 achieved the highest (or near-highest) coherence score, making it the optimal choice for representing the movie description corpus in a way that preserves semantic clarity while avoiding over-fragmentation of topics, while also not reducing the number of topics too low. The number 20 was also reasonable to accept because it lies on the plateau of value between 15 and 22.

As was described in the methodology section, it was needed to develop and train three two-tower models for Simple Retrieval, Rich Retrieval and Ranking submodels. Each submodel requires independent training, but manually selecting optimal parameters is inefficient. To improve this process, TensorFlow and Keras tuning tools were used to identify the best hyperparameters for minimizing train and validation loss and maximizing categorical accuracy. A set of 3 parameters was selected for tuning: learning rate, embedding size and batch size. The batch size was only selected to find optimal values for simply faster training on limited computing powers. Selecting optimal hyperparameter values was essential to make sure submodels, especially two-tower ones, actually learn something. However, it was also critical not to allow them to overfit.

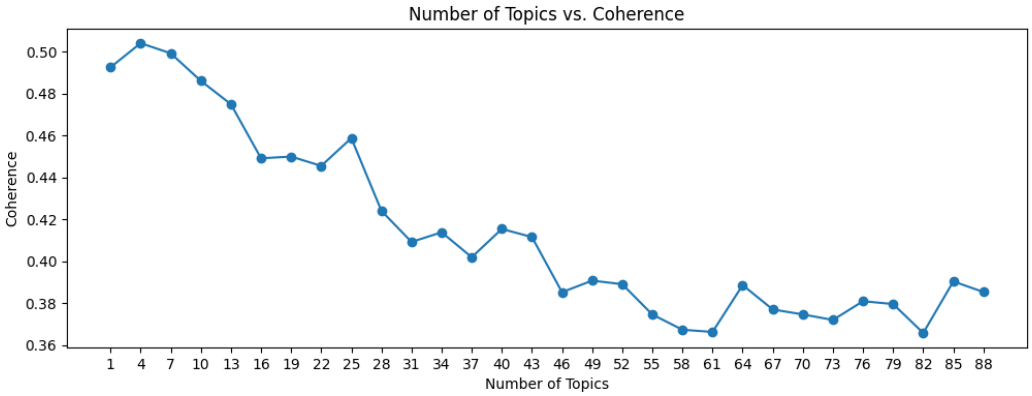


Figure 2: Topic Coherence vs. Number of Topics

Two-Tower Submodel Name	Optimal hyperparameters values
Simple Retrieval	Learning rate: 0.01 Embedding size: 64 Number of epochs before overlearning: 37
Rich Retrieval	Learning rate: 0.1 – 0.2 Embedding size:32 Number of epochs before overlearning: 4
Ranking	Learning rate: 0.1 Embedding size: 32 Number of epochs before overlearning: 20

Table 1: Optimal hyperparameter values for Two-Tower Models

Two-Tower Submodel Name	Results
Simple Retrieval	Training: top_100_categorical_accuracy: 0.2478 Validation: top_100_categorical_accuracy: 0.1163
Rich Retrieval	Training: top_100_categorical_accuracy: 0.1095 Validation: top_100_categorical_accuracy: 0.05
Ranking	Training: root_mean_squared_error: 0.8870 Validation: root_mean_squared_error: 0.8854

Table 2: Training and validation results of Two-Tower Models

Resulting values for selected parameters are shown in Table 1, and submodel training and evaluation results are shown in Table 2. The top_100_categorical_accuracy metric is a special case of the FactorizedTopK metric, which measures how often the true candidate is in the top K candidates for a given query. As was mentioned in the Prototypes Design section, this research is focused on implementing, evaluating, and comparing 3 hybrid prototypes. The first one to implement was the "SVD and LDA". The overall process for generating recommendations using this prototype consists of a few consecutive steps. First, find the top N-rated movies by the target user, sorted by the highest average rating across all users. Second, find movies similar to them by applying an item-based approach and selecting the ones with the highest similarity score. Third, apply the LDA algorithm to the selected set of movies to expand the recommendation list with films with similar summaries or descriptions. The final step was similar to all further implemented hybrid recommenders – filter only those movies that the user had not previously watched or rated and sort them by average ratings. The output result looks like a list of recommended movie IDs` with matching movie titles.

The second prototype implemented – "Simple Retrieval, Ranking, SVD and LDA" was a synthesis of the Simple Retrieval submodel, the Ranking submodel, and the previously described SVD and LDA submodels. A Simple Retrieval submodel is initially executed, receiving a specific user ID and returning a list of movie IDs that the target user will probably find interesting. The resultant list is then passed into the Ranking submodel, which subsequently produces a list of the maximum top 10 movie IDs, sorted by the ratings determined for a specific user. The next execution process is identical to the steps described for the "SVD and LDA" model, with the only difference being input movie IDs for the SVD submodel are those generated by the Retrieval and Ranking submodels.

The last developed hybrid prototype was the "Rich Retrieval and Ranking" model. The steps are similar to those described for the Simple Retrieval submodel, except that Rich Retrieval requires two input parameters: a target user ID and a timestamp value. The timestamp value is a numerical value obtained by converting the desired date to the appropriate format. In this study, timestamp values were chosen from dates near the end of September 2018.

The output forms for each hybrid prototype are similar, with only the number of recommended movies in the final resulting list varying.

Patterns of Prototypes

After successfully implementing all three hybrid models, it may be necessary to elaborate on the specific features and negative aspects of each of them. First, the "SVD and LDA" hybrid model will generate the same results for the same user every time recommendations are asked for. This problem is easily solved by passing to the SVD submodel, not just the top-N user-rated items, which are the same every time, but a random sample of a certain size from a larger set of highly rated movies. That will make recommendations more diverse and different every time users access the system. However, to evaluate the models` results, it was decided that the recommendations list should not be changed between prototype runs. The main advantage of using LDA in reviewed hybrid models is that, unlike the SVD, it does not need to be retrained every time a new item is added to the movie directory. Because each new film comes with its own description or plot summary, similar films may be found by previously trained LDA and thus appear in any user recommendation list. Because of the mentioned 'new-user' and 'new-item' problems, the SVD cannot generate a list for a user that was never seen during model training; therefore model needs to be retrained periodically every time new data arrives. The greatest advantage was gained from adopting a two-tower architecture for the Ranking and Retrieval submodels. Those may generate at least some recommendations for any user who has never been seen by the system. Furthermore, those submodels do not require as frequent retraining as the SVD submodel, because, as mentioned in the literature review chapter, the saved embeddings are fast to compute, making updates less complex and a quicker task.

In practice, the two-tower retrieval and ranking submodels can output recommendations for new users by leveraging side information (e.g., demographic or contextual features), even in the absence of explicit rating history. However, these recommendations are generally less accurate than those for experienced users, and the SVD component must be retrained periodically to fully integrate new profiles. Thus, while the hybrid setup reduces the severity of cold-start effects, it does not completely eliminate them.

5. PROTOTYPE EVALUATION

Offline testing presents several challenges that limit our ability to perform real-time evaluation. Relying solely on offline metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)—which are ultimately derived from model outputs—may not sufficiently capture predictive accuracy. Moreover, without the capacity to conduct controlled experiments with a large user base, it becomes difficult to assess user satisfaction or determine whether the recommended items are genuinely appreciated. As a result, accurately estimating precision and recall metrics remains a significant challenge.

Therefore, it was suggested to create or separate several imaginary user profiles with specific watch and rating histories. These profiles may include 'experienced' users as well as users with almost any prior system usage history. The approach involved simulating these users and having a group of respondents express their opinion on how each of the three presented prototypes coped with generating recommendations for a separate group of fictitious users. The target group was offered a questionnaire in which they expressed their overall satisfaction with certain models numerically and indicated the number of movies or titles from the offered recommendation list, which they believed the users would enjoy the most.

From the MovieLens dataset description, we know that any user had left at least 20 ratings from a selected set of movies. To increase the diversity of the user profiles chosen for testing and evaluation, it was decided to select five users from among 610: User 1, User 2, and User 3 are considered the most experienced users, with between 450 and 550 reviews, while Users 4 and 5 are considered the least experienced users, with only 20 ratings recorded.

A preliminary survey was carried out to assess the recommender systems. Through Google Forms, eleven respondents were asked to rate the models' overall performance (on a scale from 1 to 10, with 1 indicating disastrous performance with completely irrelevant recommendations, and 10 – top-tier performance, all recommendations are up to user taste and anticipations and provide the number (3,7,10, etc.) of well-predicted films among all the recommendations made by specific methods. Profiles that contained the viewing and rating history of each of the five test users were provided. Survey participants were required to simulate the target users based on a specified profile and decide whether they liked recommended movies from the provided generated lists or not.

Most of the respondents were graduate students, ranging in age from 23 to 27. Their knowledge of movies and the film domain varied greatly; some had little experience with watching and analysing movies, while others were highly knowledgeable in this area. As a result, those groups had the most difficulty assessing the recommender model results because they either knew too little about it or were overanalysing and complicating their conclusions, causing a possible bias.

After the review gathering process was completed, each model's precision and average rating could be computed. To determine the precision score, the average ratio of how frequently movies from generated lists were liked by a user was calculated. The survey shows the preliminary results presented in Table 3.

Prototype	Precision	Score
Simple Retrieval-Ranking, SVD and LDA	57%	6.72
SVD and LDA	53%	6.36
Rich Retrieval-Ranking	39%	4.36

Table 3: Preliminary Survey Results of Prototypes Performance

6. DISCUSSION

This study designed, analysed, and compared three hybrid movie recommender systems integrating classical content-based and collaborative filtering techniques alongside a recent two-tower neural network model. All the prototypes demonstrate the potential to generate movie recommendations for selected users.

According to the previous chapter's questionnaire results, the "Simple Retrieval-Ranking, SVD and LDA" model emerged as the most effective in providing suitable, accurate, and curated recommendations. Both average score and precision turned out to be the best. The hybrid approach and continuous refinement of intermediate results proved their adoption and yielded excellent results.

To conclude the results and determine whether there is a statistical difference between the two models, a two-sample t-test was completed. To ensure that we can apply the t-test to the data

we had, it was first confirmed that the two samples of precision data have a normal distribution using the Shapiro-Wilk test, and the variances are not too different. The received p-value of 0.31 is bigger than the significance level of 0.05, meaning that we should accept the hypothesis that the mean precisions of the first two models are equal.

Interestingly, prototypes' precision was the same for experienced users, but for users who had watched the fewest movies, the difference between top prototypes was much greater, more than 10%. However, the results of the Rich Retrieval Ranking prototype were deemed unsatisfactory. There are several possible explanations. During the review-gathering process, a few complications arose. In addition to the previously mentioned different movie backgrounds of respondents, some people expressed their views on the fact that the movie lists generated by the "Rich Retrieval-Ranking" prototype included very obscure titles that they had no idea how to evaluate. These issues could have significantly impacted the results. A larger-scale evaluation or a change in the evaluation process may be needed to confirm the results or identify any misconceptions about them.

While the evaluation of recommendation accuracy and user satisfaction was based solely on survey responses from participants who assessed model outputs for only five test user profiles, it is essential to acknowledge the inherent limitations of this assessment approach. As a result, the obtained findings may not be entirely conclusive, and further testing and analysis may yield unexpected results, necessitating a more comprehensive evaluation of the developed models.

The second prototype, "Simple Retrieval-Ranking, SVD and LDA", yielded the most impactful results. Each submodel was sufficiently trained, allowing their combination to generate valuable movie recommendations. Moreover, the system shows promise for further enhancement with the incorporation of additional training data.

Individual testing suggested that the "Rich Retrieval-Ranking" prototype tends to overtrain, potentially due to the inclusion of the timestamp feature as one of its embeddings. Further investigation is necessary to confirm whether this factor is the root cause.

7. FUTURE WORK

The effectiveness of generating comprehensive and high-quality recommendations with the proposed prototypes is significantly constrained by the size of the initially selected dataset. Achieving higher accuracy and improved user feedback for each model would likely require a dataset several orders of magnitude larger, thereby increasing the volume of data, the number of users, and the diversity of movies.

A key avenue for improvement involves refining the two-tower model architecture. Future iterations may incorporate additional features beyond the existing timestamp and movie titles, such as user age, movie genre, and other relevant attributes. Moreover, the removal of the timestamp feature is also under consideration, though further experimentation and evaluation will be necessary to assess its impact.
Additionally, new approaches to model evaluation could be explored. One potential direction is leveraging AI-driven chatbot agents to assess the quality of the generated recommendation lists. By equipping a GPT-based agent with historical user rating data and provided recommendations, it may be possible to simulate user preferences and evaluate how well the recommended movies align with a target user's interests.

By changing the model architecture, training and test dataset sizes and evaluation approach, it might be possible to improve recommendation accuracy and enhance models' speed, versatility and flexibility.

8. CONCLUSION

This research contributes to the recommender systems domain by developing and evaluating three distinct recommendation models, two of which performed well based on precision and average rating metrics. The work highlights the effectiveness of selected hybrid modelling strategies for personalised recommendations and provides a comparative analysis that may inform future model selection in similar contexts. Additionally, the project introduces practical insights on balancing accuracy and diversity in recommendation outputs. Overall, it offers a framework and empirical findings that can support continued improvement in recommender system design.

In conclusion, the results obtained and the overall performance of "Simple Retrieval-Ranking, SVD and LDA" with "SVD and LDA" hybrid models proved their potential. However, there remains significant scope for further refinement and improvement.

9. REFERENCES

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3, 993-1022.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4), 331-370. https://doi.org/10.1023/A:1021240730564

Chang, S. Y., Wu, H. C., Yan, K., Chen, X., Huang, S. C. H., & Wu, Y. (2023). Personalized multimedia recommendation systems using higher-order tensor singular-value-decomposition. *IEEE Transactions on Broadcasting*, 70(1), 148-160. https://doi.org/10.1109/TBC.2023.3278111

Cremonesi, P., Koren, Y., & Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, 39-46. https://doi.org/10.1145/1864708.1864721

Falk, K. (2019). *Practical recommender systems*. Simon and Schuster.

Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. *ACM transactions on interactive intelligent systems (tiis)*, 5(4), 1-19. https://doi.org/10.1145/2827872

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, 16(3), 261-273. https://doi.org/10.1016/j.eij.2015.06.005

Javed, U., Shaukat, K., Hameed, I. A., Iqbal, F., Alam, T. M., & Luo, S. (2021). A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning*, 16(3), 274-306. https://www.learntechlib.org/p/219036/

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78(11), 15169-15211. https://doi.org/10.1007/s11042-018-6894-4

Klema, V., & Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2), 164-176. https://doi.org/10.1109/TAC.1980.1102314

Konstan, J. A., & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User modeling and user-adapted interaction*, 22(1), 101-123. https://doi.org/10.1007/s11257-011-9112-x

Lee, W. M., & Cho, Y. S. (2023). A Flexible Two-Tower Model for Item Cold-Start Recommendation. *IEEE Access*, 11, 146194-146207. https://doi.org/10.1109/ACCESS.2023.3346918

Li, L., Zhang, Y., & Chen, L. (2023). Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, 41(4), 1-26. https://doi.org/10.1145/3580488

Papadakis, H., Papagrigoriou, A., Panagiotakis, C., Kosmas, E., & Fragopoulou, P. (2022). Collaborative filtering recommender systems taxonomy. *Knowledge and Information Systems*, 64(1), 35-74. https://doi.org/10.1007/s10115-021-01628-7

Thannimalai, V., & Zhang, L. (2021). A content based and collaborative filtering recommender system. In *2021 International Conference on Machine Learning and Cybernetics (ICMLC)*, IEEE, 1-7. Https://doi.org/10.1109/ICMLC54886.2021.9737238

Thorat, P. B., Goudar, R. M., & Barve, S. (2015). Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, 110(4), 31-36. https://doi.org/10.5120/19308-0760

Varsha, M. (2024). Two tower recommendation system. *Medium*. Retrieved August 2025 from https://medium.com/@varshamoturi3/two-tower-recommendation-system-d2da761fcbce

Wang, Y., Xiong, F., Han, Z., Song, Q., Zhan, K., & Wang, B. (2025). Unleashing the Potential of Two-Tower Models: Diffusion-Based Cross-Interaction for Large-Scale Matching. In *Proceedings of the ACM on Web Conference 2025*, 304-312. https://doi.org/10.1145/3696410.3714829

Wortz, J., & Totten, J. (2023). Scaling deep retrieval with TensorFlow Recommenders and Vertex AI Matching Engine. *Google Cloud*. Retrieved August 2025 from. https://cloud.google.com/blog/products/ai-machine-learning/scaling-deep-retrieval-tensorflow-two-towers-architecture

Yang, J., Yi, X., Zhiyuan Cheng, D., Hong, L., Li, Y., Xiaoming Wang, S., ... & Chi, E. H.

(2020). Mixed negative sampling for learning two-tower neural networks in recommendations. In *Companion proceedings of the web conference 2020*, 441-447. https://doi.org/10.1145/3366424.3386195

Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., ... & Li, Q. (2024). Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, 36(11), 6889-6907. https://doi.org/10.1109/TKDE.2024.3392335

Evaluating AWS vs. Azure for Generative AI in Healthcare: A Comparative Analysis Using the NIST CSF 2.0 Maturity Model

Eli Taylor
tayloreli@cityuniversity.edu

Scott Zhou
zhouscott1@cityu.edu

Juan Carlos Garcia
garciajuancarlos1@cityuniversity.edu
City University of Seattle and Universidad Panamericana (Mexico City, Mexico)

Brittney Cherry
cherryb@cityuniversity.edu

Sam Chung
chungsam@cityu.edu

City University of Seattle
Seattle, WA 98121

Abstract

The rapid adoption of generative artificial intelligence (GenAI) technologies, healthcare organizations aiming to leverage these advancements often need help scaling their personnel and infrastructure. Amazon Web Services (AWS) and Microsoft Azure are leading cloud service providers. This study aims to analyze and compare the GenAI offerings from AWS and Azure. AWS offers robust GenAI tools like SageMaker, while Azure counters with Azure Machine-Learning. We comprehensively review their capabilities and potential to enhance organizational maturity in operational efficiency and innovation within the health insurance sector. Utilizing the maturity model within the NIST Cybersecurity Framework (CSF) 2.0, we will evaluate how generative AI solutions from these cloud platforms can contribute to improving healthcare organizational maturity. Our methodology encompasses a framework proposal for analyzing GenAI technologies, a review, and a comparative analysis between AWS and Microsoft. This ensures a robust integration with NIST CSF 2.0, specifically addressing healthcare organizations. We perform an in-depth examination of case studies, industry reports, and existing literature to provide a nuanced understanding of each platform's strengths and weaknesses. We will also consider cost, ease of use, scalability, and integration with existing healthcare systems. The research question for this paper is: how do AWS and Azure capabilities differ in GenAI capabilities and features, and how can these differences impact organizational maturity in the healthcare industry, as defined by the NIST 2.0 framework. Our analysis has shown that both AWS and Azure offer foundational solutions capable of supporting the deployment of GenAI in health insurance organizations, each with distinct strengths.

Keywords: Microsoft Azure, Amazon Web Services, Generative AI, National Institute of Standards and Technology, Cybersecurity Framework, Health Care

Recommended Citation: Taylor, E., Zhou, S., Garcia, J., Cherry, B., Chung, S., (2026). Evaluating AWS vs. Azure for Generative AI in Healthcare: A Comparative Analysis Using the NIST CSF 2.0 Maturity Model. *Journal of Information Systems Applied Research and Analytics*, v19(n?) pp 26-37. DOI# https://doi.org/10.62273/SHFC3950

Evaluating AWS vs. Azure for Generative AI in Healthcare: A Comparative Analysis Using the NIST CSF 2.0 Maturity Model

Eli Taylor, Scott Zhou, Juan Carlos Garcia, Brittney Cherry and Sam Chung

1. INTRODUCTION

In healthcare, data-driven decision-making is essential for enhancing patient care and improving clinical outcomes. The healthcare stakeholder approach to GenAI adoption involves distinct objectives and contexts. Providers emphasize diagnosis and patient communication, while payers (Health Care Organizations) claim efficiency, fraud detection, and cost reduction (Accenture, 2023; Johnson et al, 2021). Regulators prioritize standardized compliance like the Health Insurance Portability and Accountability Act (HIPAA), where violations can result in several financial and reputational risks (Nancy & Kumar, 2023). These differences show the need for a GenAI implementation framework employing NIST CSF 2.0 for assessing their readiness across Govern, Protect, Detect, Response, and Recovery functions.

To stay competitive, organizations must embrace digital transformation, including cloud migration and advanced analytics (García-Peñalvo & Vázquez-Ingelmo, 2023). The potential of generative AI (GenAI) to revolutionize healthcare is vast, with applications ranging from creating medication instructions and marketing content to developing AI-driven healthcare delivery methods like chatbots for mental health counseling (Chui et al., 2023a; Kanbach et al., 2024). Given the complexities of implementing GenAI in a regulated environment, adopting a maturity model offers organizations a systematic framework to guide GenAI deployment, ensuring that innovation aligns with operational efficiency, security, and compliance requirements.

Healthcare organizations can make a foundational effort to enhance their technical capabilities and advance cybersecurity maturity by leveraging cloud platforms such as AWS and Azure. Both companies offer cloud technologies and related services to fulfill cybersecurity regulatory services for data care and operational continuity. This paper explores a framework application study on AWS, and Azure technologies supporting the implementation of GenAI in healthcare, with a focus on how these platforms align with the NIST CSF 2.0 maturity model to drive organizational growth and resilience.

Healthcare Industry GenAI Challenges

GenAI refers to machine learning methods that generate new content such as text, images, or speech in response to human inputs (Cao et al., 2018). In healthcare, GenAI can support diagnosis, enhance patient interactions, and automate documentation. However, adoption faces challenges including cost of implementation, PHI protection, explainability, and ethical concerns (Reznikov, 2024; Chui et al., 2023b). Workforce readiness and data governance further complicate deployment (Hennrich et al., 2024). Despite these obstacles, benefits include earlier disease detection (Zhang & Boulos, 2023), reduced claims costs (Accenture, 2023), and efficiency gains in administrative workflows (Berlin et al., 1997). Cloud services such as AWS and Azure offer scalable infrastructure to address these needs, but evaluating them requires a framework that integrates cybersecurity and compliance maturity — provided here by the NIST CSF 2.0. Frameworks like the NIST CSF 2.0 and the NIST AI Risk Management Framework provide structured approaches to aligning AI adoption with security and compliance needs (Renkema, 2023; Manek et al., 2024).

The use of GenAI in this context is governed by more than just potential efficiency gains. The industry is subject to a stringent regulatory environment designed to protect sensitive Protected Health Information (PHI). While compliance with HIPAA is a baseline requirement, the concerns extend further. A superficial approach focused solely on avoiding substantial fines" overlooks the broader and more critical issues of reputational, cybersecurity risks, data leakage, the erosion of member trust, and ethical issues (Chen & Esmaeilzadeh, 2024). The deployment of AI models that are biased, non-transparent, or insecure can have profound negative consequences, leading to inequitable outcomes and legal challenges.

Project Overview
This project aims to analyze and compare the GenAI offerings of Amazon Web Services (AWS) and Microsoft Azure, focusing on their potential to enhance operational efficiency, innovation capacity, and cybersecurity maturity in health insurance organizations. By examining cost, ease of use, scalability, security features, and integration capabilities, this study provides healthcare decision-makers (Chief Executive Officer, Chief Operating Officer, Chief Medical Officer, Chief Nursing Officer, Department Heads, Chief Information Officer and Chief Information Security Officer) with insights into advancing their organizations' technical maturity while maintaining robust security and governance.

Program Mission
This project seeks to deliver insights for healthcare decision-makers (Chief Operating Officer, Chief Medical Officer, Chief Nursing Officer, Department Heads, Chief Information Officer and Chief Information Security Officer) seeking to advance their organizations' technical and cybersecurity maturity by implementing GenAI technologies. This project will assess how AWS and Azure support each stage of the maturity model, from governance and risk management to protection, detection, response, and recovery. By comparing these platforms within the maturity model framework, the study seeks to identify the optimal path for healthcare organizations to harness the power of GenAI while achieving higher levels of organizational maturity and operational excellence.

External and Internal Influencers
Multiple external and internal factors influence the successful implementation and adoption of generative AI technologies within healthcare organizations. Understanding these factors is critical for healthcare decision-makers (Chief Executive Officer, Chief Operating Officer, Chief Medical Officer, Chief Nursing Officer, Department Head, Chief Information Officer and Chief Information Security Officer) aiming to enhance their organizational maturity through advanced AI and cloud solutions.

External Influences
There are multiple external factors for data care that a health insurance company must consider when deciding to deploy and use generative AI, like clinical practice, medical imaging and data augmentation, drug discovery and biomedical research addressed to HIPAA supervision, European Research Council (ERC) and National Science Foundation (NSF) (Rabbani et al, 2025) . First, compliance with the Health Insurance Portability and Accountability Act (HIPAA) is crucial, as violations can result in substantial fines (Nancy & Kumar, 2023). HIPAA sets standards for data security and protection a health insurance company must adhere to. Aside from HIPAA, implementing GenAI can come with the potential benefits of being the first company to create an innovative product or service. This may include reduced expenses and increased revenues (Anand, 2024).

The health insurance company must also determine what type of potential grants or outside opportunities for GenAI financial support are available. If the company can secure grants and partnerships that reduce the company's initial investment, then the company is more likely to consider GenAI. However, if the company must take on all the costs without any outside funding, the potential for innovation into GenAI is reduced (Hennrich et al., 2024)

Internal Influencers
Internal culture is one of the strongest determinants of whether GenAI will be utilized within a health insurance company. Suppose the company culture is generally opposed to changes and innovation. In that case, it will be much harder for the company to get buy-in from employees, and the success of GenAI within the company is considerably reduced (Hennrich et al., 2024).

The skillset of the internal workforce should be considered, too. If the company's employees lack skills in GenAI or cloud computing, the company will find deploying these innovative solutions costly and at high-risk. Potential solutions include hiring employees with these skills, training current employees, or having technology that is underutilized. Therefore, the cost to hire or train a workforce in GenAI and cloud computing increases significantly and may be more than the company is willing to spend (Hennrich et al., 2024).

2. BACKGROUND

Generative AI
Generative AI, or GenAI, refers to machine learning methods that extract intent from human requests and generate relevant content in response (Cao et al., 2018). Applications include computer vision, text generation, music composition, and speech synthesis, supported by deep neural networks trained on massive datasets using advanced processing power

(Dasgupta et al., 2023; García-Peñalvo & Vázquez-Ingelmo, 2023). Modern GenAI architectures include Generative Adversarial Networks (GANs) for computer vision and Generative Pre-Trained Transformers (GPTs) for natural language processing (Reznikov, 2024).

Adopting cloud computing has made GenAI more accessible by providing scalable, cost-effective infrastructure. While challenges like high implementation costs and complex integration remain, cloud platforms and open-source models offer scalable resources and simplified customization (Lu et al., 2024). The emergence of Large Models as a Service, such as ChatGPT, has further democratized access to GenAI technology.

Challenges for Healthcare Organizations

While GenAI offers promising opportunities for health insurance, common issues across industries include the cost of implementation, complex integration, lack of skilled professionals, and concerns about data privacy and security (Reznikov, 2024). Organizations must carefully evaluate the costs of cloud migration, infrastructure redesign, and workforce development against the potential return on investment (Hennrich et al., 2024).

Health insurance organizations face additional industry-specific challenges, such as data privacy, explainability, ethical concerns, and fault tolerance (Cao et al., 2018). Strict protected health information (PHI) regulations require well-configured cloud infrastructure and robust security governance measures (Chui et al., 2023b). Applications that directly interface with patients may have little to no fault tolerance, requiring rigorous testing and monitoring for safety, accuracy, and bias prevention (Cao et al., 2018). Finally, ensuring explainability and transparency in GenAI is crucial, particularly in high-risk settings, and may require human oversight for validation.

Benefits for Healthcare Organizations

Despite these challenges, GenAI has the potential to enhance patient care significantly. By enabling physicians to diagnose diseases earlier and with greater accuracy, GenAI allows physicians to focus on complex issues and enables more effective communication with patients (Zhang & Boulos, 2023).

For GenAI to benefit patients and physicians, healthcare companies must justify the implementation costs. Early disease detection through GenAI reduces patient care costs and decreases lawsuits from incorrect or missed diagnoses (Zhang & Boulos, 2023). GenAI can also help reduce the utilization of high-cost services in non-emergency situations by identifying patient issues that can wait until regular office hours (Travers, 2003).

Contrary to the belief that insurance companies might lose revenue due to GenAI, the reality is that there are not enough medical providers to meet the current demand. GenAI allows providers to bill insurance companies while reducing the cost of care, benefiting providers and insurers (Shryock, 2022). Moreover, GenAI can automate administrative tasks such as scheduling, updating lab results, and communicating with patients, lowering costs and RVU (Relative Value Unit) expenses for medical providers and insurance companies (Berlin et al., 1997).

Ultimately, by decreasing unnecessary emergency room visits, lawsuits, late diagnoses, and high administrative costs, GenAI can make healthcare more affordable, enhancing the quality of care for everyone, provided its use is maximized across the industry.

Strategic Goals and Objectives

Strategic goals for the health insurance industry include reducing costs, improving claims processing efficiency, and enhancing quality controls. Health insurers can leverage GenAI and cloud services like AWS or Azure to achieve these objectives. According to McKinsey, healthcare technologies could reduce healthcare spending by 8-12% in 14 countries, with 30% of these savings benefiting insurers through reduced claims and improved risk management (Nathella, 2024). Additionally, Accenture found that AI-driven technologies could cut claims processing costs by 20-25% through automation and enhanced fraud detection (Accenture, 2023).

While the cost and quality benefits are clear, GenAI and cloud services like Azure or AWS also allow health insurance companies to customize rates and tailor services to each customer's current and predicted health status. The outcome is a population health approach, where the services provided are customized (Johnson et al., 2021). Through partnerships between insurance companies and their customers, the insurance company and the overall health of the community benefit.

3. LITERATURE REVIEW

Comparative Analysis: AWS vs. Azure

GenAI in AWS

With the largest market share as a cloud service provider, AWS offers extensive, customizable services designed to support machine learning infrastructure in health insurance organizations. One of AWS's critical offerings in this area is Amazon Bedrock, a service that integrates several leading Large Language Models (LLMs), (*Foundation Model API Service – Amazon Bedrock,* n.d.).

Amazon Bedrock allows organizations to create tailored LLM instances with proprietary datasets within a secure, encrypted environment. To achieve this, Bedrock creates a copy of the selected LLMs. It allows the addition of specialized data sets to enrich the Foundation Model (FM) over an encrypted environment that does not refeed the original FM. Retrieval-Augmented Generation (RAG), which is well supported as well, allowing users to personalize models by adding curated documents to increase response relevance and accuracy in domain-specific settings like healthcare (*Build Generative AI Applications with Foundation Models - Amazon Bedrock - AWS*, 2024). This makes it well-suited for microservices architectures, allowing health insurance companies to develop tailored AI-driven solutions.

GenAI in Azure

Azure's GenAI offerings include a comprehensive suite of products and services that can add value within the health insurance industry. These services include Azure Machine Learning, an end-to-end platform for building, training, and deploying machine learning models, and development tools like Azure AI Studio and Azure Databricks for collaborative data science and analytics (*AI and Machine Learning - Azure Services,* n.d.).

Health insurance companies seeking to streamline data analysis, enhance customer service, and optimize operational efficiencies will have numerous specialized services. For example, a health insurance company developing a customer service chatbot can streamline development with Azure AI Bot Service or Azure Health Bot to deliver customer interactions (*AI and Machine Learning - Azure Services,* n.d.).

Azure's predictive analytics capabilities offer significant benefits for health insurers. Through Azure's scalable cloud infrastructure, insurers can analyze datasets, identify patterns, predict trends, and make informed decisions without requiring extensive in-house resources.

AWS vs. Azure Considerations

As leading cloud service providers, AWS and Azure provide comprehensive solutions for health insurers ready to migrate to the cloud and leverage the full capabilities of GenAI tools and services. The decision to use one platform over the other should be based on the specific needs of the organization and a thorough understanding of the strengths and limitations of each platform.

One primary consideration for companies hesitant about deploying new infrastructure or developing in-house expertise is the interoperability with their current technology stack. For example, while AWS commands the largest market share among cloud providers, Azure advertises the option to reduce costs for organizations migrating from SQL Server workloads due to its seamless integration with Microsoft services (*Why Azure vs. AWS,* 2022). Organizations already familiar with Microsoft products may find the Azure environment more comfortable and cost-effective.

Conversely, for organizations aiming to connect globally or requiring extensive data center availability, AWS may offer an advantage with its broader geographic coverage, more comprehensive service catalog, and highly customizable options (*AWS vs Azure: The Ultimate Cloud Face-Off,* 2023). This makes AWS particularly appealing for companies that need to scale quickly or require specialized cloud services.

Well-Architected Framework: AWS vs Azure

AWS and Azure have Well-Architected Framework offerings to provide governance. These frameworks consist of pillars as guiding principles and various evaluation and scoring tools to measure the effectiveness of infrastructure in practice. Both AWS and Azure tools share five pillars of operational excellence, security, reliability, performance efficiency, and cost optimization, while AWS adds a sixth pillar of sustainability (*AWS Well-Architected - Build secure, efficient cloud applications,* n.d.; *Microsoft Azure Well-Architected Framework,* 2023). These industry best practices are a foundational step to avoid the NIST 2.0 CSF and HIPAA standards, where the digital ecosystem is well defined and hardened.

Cybersecurity Framework NIST 2.0

The Security Framework referenced in this work is the CSF 2.0, published by the National Institute of Standards and Technology on February 26th, 2024 (*The NIST Cybersecurity Framework (CSF) 2.0*, 2024). Synthesized on six functions and 106 controls defined as follows:

1. Govern, 6 subcategories with 31 controls
2. Identify, 3 subcategories with 21 controls
3. Protect, 5 subcategories with 22 controls
4. Detect, 2 subcategories with 11 controls
5. Respond, 4 subcategories with 13 controls
6. Recover, 2 subcategories with 8 controls

Furthermore, for a brief understanding, we will describe briefly the CSF 2.0 controls with healthcare-specific examples:

1. **Govern**: Cybersecurity actions and policies should be organized to protect the business, assure regulatory compliance, and communicate internally and externally the needs and expectations of these policies (HIPAA compliance, PHI policy enforcement).
2. **Identify**: This point helps determine the organization's current cybersecurity risks and prioritize efforts according to business risk management (Mapping Electronic Health Records and claims assets).
3. **Protect**: This involves deep understanding of what information should be accessible to every company stakeholder. In the same way, building processes to recognize and respond to cyberattacks and suspicious activity is crucial to protect the organization. (Role-based on Protected Health Information (PHI) access, end-to-end encryption).
4. **Detect**: This point implicates understanding how to identify a cybersecurity incident by defining typical digital behavior and what is not (Anomalous data claims monitoring).
5. **Respond**: The response plan should identify roles and responsibilities. Organizations can begin by identifying the abilities, skills, and resources needed to respond to a cybersecurity incident (Incident response for breaches, fraud attempts, and response workflows).
6. **Recover**: This section defines the roles and responsibilities for recovering data inside and outside the organization. This involves assessing the health of the backup data schema for the restoring process according to organizational needs and resources (Patient record restoration after cyber-attack).

Table 1 summarizes how AWS and Azure capabilities align with each NIST CSF 2.0 function in the healthcare context. Furthermore, these distinctions suggest that platform selection should be influenced not only by technical capabilities but also by organizational maturity and ecosystem aligned with the HPI compliance and resiliency. The results indicate that AWS and Azure are both capable platforms for advancing GenAI adoption in healthcare. However, their value differs by organizational priority. AWS may be more attractive for organizations requiring extensive global coverage and high configurability, while Azure provides advantages for healthcare organizations prioritizing integration with Microsoft-based compliance and governance tools. These findings support the argument that healthcare decision-makers (Chief Operating Officer, Chief Medical Officer, Chief Nursing Officer, Chief Information Officer and Chief Information Security Officer) should align their cloud platform choice not only with current technical needs but also with their position along the NIST CSF maturity continuum oriented to HIPAA compliance.

NIST	Healthcare Requirement	AWS Capability	Azure Capability	Notes / Trade-offs
Govern	HIPAA compliance, policy integration	AWS Artifact, compliance templates	Microsoft Purview, Compliance Manager	Azure integrates well for Microsoft environments; AWS offers broader templates
Identify	Risk assessment, asset visibility	AWS Security Hub	Azure Security Center	Both automate asset/risk identification; Azure stronger in Microsoft-linked ecosystems
Protect	PHI encryption, secure access	AWS KMS, IAM, VPCs	Azure Key Vault, Conditional Access	AWS excels in encryption customization; Azure leverages Microsoft identity
Detect	Anomaly and breach monitoring	AWS GuardDuty, CloudTrail	Azure Sentinel, Log Analytics	Comparable; Azure integrates SIEM, AWS broader feeds
Respond	Incident readiness, automation	AWS Incident Manager	Azure SOAR workflows	Both enable automated playbooks; preference depends on workflow design
Recover	Backup, resilience	AWS Backup, global coverage	Azure Site Recovery, Backup Vault	

Table 1: AWS vs Azure capabilities against NIST CSF 2.0 functions with healthcare requirements.

The comparative analysis shows that healthcare organizations must align platform selection not only with technical capabilities but also with organizational maturity. AWS is helpful for organizations requiring scalability and

configurability, while Azure is ideal for those already integrated with Microsoft systems. The conceptual rubric provides decision-makers with a structured, healthcare-specific evaluation method, addressing the reviewers' critique that prior drafts resembled a vendor whitepaper.

Building a Cybersecurity Maturity Model in the Era of Artificial Intelligence
This research demonstrates applying a cybersecurity maturity model tailored for the AI era using the NIST CSF 2.0 framework. It addresses the challenges and opportunities caused by AI, including GenAI, from a cybersecurity perspective. Health insurance organizations can use established frameworks like NIST to manage emerging risks and integrate GenAI technologies to improve cybersecurity. We highlight a framework for integrating GenAI technologies within an organization's broader cybersecurity strategy (Renkema, 2023).

Generative Artificial Intelligence profile by NIST
The paper provides a comprehensive profile of generative AI, covering risks and benefits, supported by the NIST AI Risk Management Framework. It discusses data security, ethical considerations, and implementation challenges, emphasizing the importance of applying the NIST maturity model to GenAI (NIST, 2024).

Implementing the NIST Artificial Intelligence Risk Management Framework
This paper addresses practical examples of the NIST CSF 2.0 AI Risk Management Framework, guiding organizations in managing AI-related risks. It covers secure, compliant, and ethical AI deployment, including resources to implement a maturity model for GenAI technologies and insights into achieving higher maturity levels (Manek et al., 2024).

4. METHODOLOGY

Metrics Used to Measure Outcomes
A set of robust metrics is required to measure the impact of GenAI in healthcare, especially within cloud computing. These metrics assess operational deficiency, innovation capacity, cost management, and organizational maturity level as defined by the NIST 2.0 framework (NIST, 2024).

Operational efficiency metrics evaluate how GenAI streamlines healthcare operations. Key metrics include:

- Time to Insight, tracks the time spent from data ingestion to generating insights (García-Peñalvo & Vázquez-Ingelmo, 2023).
- Resource Utilization, measures computational resource use, aiming for higher utilization with lower costs.
- Task Automation Rate, assesses the percentage of routine tasks completed by GenAI to indicate efficiency change by automation.
- Response time in Patient Interaction, tracks the speed of AI-powered tools to respond to patient inquiries. Reduced times indicate service efficiency (Cao et al., 2018).

Innovation Capacity Metrics show the healthcare industry's capability to innovate with GenAI technologies. These include:

- The number of New AI-driven Applications indicates how many new GenAI apps were developed and deployed.
- The adoption rate of AI Solutions measures how many departments have embraced these technologies, with a higher adoption rate indicating successful integration.
- Idea-to-implementation Cycle Time, evaluates the time taken from concept to deployment of the new GenAI app; shorter cycles show greater agility (Lu et al., 2024).

Cost Efficiency Metrics are essential to evaluating the financial impact of GenAI technologies. Key metrics are:

- Total cost of ownership (TCO) assesses all costs related to deploying the GenAI technologies, with lower TCO indicating cost efficiency.
- Return on Investment (ROI), measures the financial return from GenAI technologies relative to their costs; higher ROI indicates successful outcomes (Kanbach et al., 2023).

Organizational maturity metrics align with the NIST CSF 2.0 framework, measuring progression in cybersecurity maturity.

- Maturity level assessment evaluates the organization's maturity level across all components from the NIST CSF 2.0 framework; higher levels indicate better cybersecurity practices.
- Risk management efficiency**,** measures the effectiveness of strategies for managing risks associated with GenAI.
- Incident Response Time, tracks the time taken to detect, respond, and recover from cybersecurity incidents involving GenAI. Shorter response times indicate strong resilience (Eiras et al., 2024).

Limitations
While this study applied a structured analysis over the NIST CSF 2.0 rubric, AWS and Microsoft cloud providers, it did not include empirical testing over AWS or Azure healthcare environments. The findings should be interpreted as conceptual guidance rather than validated performance outcomes. Future research should incorporate empirical testing, organizational surveys, and performance testing to evaluate the framework's applicability and extend it to patient benefits, cost efficiency, and ethical considerations for healthcare organizations.

5. RESULTS

Specific Initiatives and Timelines for Implementation
A structured approach can help align the NIST CSF 2.0 maturity model to GenAI technologies in health insurance organizations. The process will be divided into several phases: assessment, infrastructure enhancement, deployment, and improvement.

Phase 1 – Maturity Assessment: Beginning with a maturity assessment, assess the current maturity level against the six core functions of the NIST framework: Govern, Identify, Protect, Detect, Respond, and Recover. The assessment focuses on cybersecurity practices and the status of GenAI implementation, providing a detailed report outlining the current maturity level, gaps, and areas for improvement. This phase could last between zero and three months (Renkema, 2023).

Phase 2 – Infrastructure Enhancement and Training: This phase focuses on enhancing security infrastructure and workforce skills to support GenAI deployment. This covers encryption protocols, access controls, threat detection systems, and data protection measures. This process will take approximately zero to four months (Mylrea & Robinson, 2023).

Phase 3 – GenAI Deployment and Integration: In this phase, GenAI deployment and integration transition from planning to implementation after testing in a controlled environment. This stage focuses on automating routine tasks, improving patient interactions with AI chatbots, and implementing GenAI within existing healthcare systems. This phase will take zero to three months (Chui et al., 2023b).

Phase 4 – Continuous Improvement and Monitoring: This phase begins with establishing continuous monitoring systems. These systems mainly track the performance of GenAI solutions using cybersecurity measures such as regular audits, real-time monitoring, and performance evaluation Key Performance Indicators (KPIs). This phase will take zero to five months (Vakkuri et al., 2021).

Phase 5 – Long-term Optimization and Expansion: This phase involves optimizing and expanding GenAI solutions within the organization. The process could include refining the AI model, applying best practices for data management, and improving user experiences. In addition, continuous improvement is needed to address new threats and ensure long-term resilience. This phase lasts between zero and five months (Lu et al., 2024).

6. DISCUSSION

Comparative Analysis
We structure the comparison of AWS and Azure services around the six core functions of the NIST CSF 2.0. This approach focuses on how each platform supports health insurance organizations in progressing through the critical stages of cybersecurity maturity while implementing GenAI solutions.

AWS and Azure effectively support health insurance organizations over NIST CSF 2.0 maturity model, offering robust solutions for governance, risk identification, data protection, threat detection, incident response, and recovery. However, there are distinctions worth noting. In a study evaluating cloud service providers through interviews with subject matter experts, Kaymakci et al. (2022) found that Azure ML Studio surpassed AWS SageMaker regarding performance, reliability, and cloud management, while AWS provided better flexibility and cost-effectiveness. Notably, both providers were ranked the same regarding security features, providing comprehensive security measures and tools that enhance cybersecurity posture and ensure regulatory compliance (Kaymakci et al., 2022).

Evaluation of Outcomes
Implementing GenAI in healthcare presents a transformative opportunity to enhance patient care through improved operational efficiency, innovation capacity, and cost management. Health insurance companies can advance their technical and cybersecurity maturity by leveraging the features available through cloud

platforms like AWS and Azure. The successful deployment of GenAI, supported by a robust NIST CSF 2.0 framework, ensures that organizations can address the complex regulatory environment of healthcare, maintain compliance, and protect patient data.

Specific metrics aligned with the NIST CSF 2.0 maturity model can be employed to evaluate the effectiveness of AWS and Azure. For example, operational efficiency can be measured using Time to Insight" and "Task Automation Rate." At the same time, innovation capacity can be assessed through the "Number of New AI-Driven Applications" and "Adoption Rate of AI Solutions." Financial impact metrics like "Total Cost of Ownership" (TCO) and "Return on Investment" (ROI) can help evaluate the cost-effectiveness of the deployed GenAI solutions. By integrating these metrics with the NIST CSF 2.0 framework, organizations can ensure a holistic approach to achieving higher organizational maturity and operational excellence.
Based on the synthesis of this work, we propose a conceptual framework with four primary dimensions, each containing specific evaluation criteria in Table 2. This framework is designed to be a practical tool for health insurance decision-makers to conduct a systematic and comprehensive analysis of cloud GenAI platforms for healthcare organizations.

Dimension	Evaluation Criterion	Description	Justification
1. Technical Performance & Capabilities	Model Diversity & Quality	Access to a range of high-quality foundation models (proprietary and third-party).	Avoids vendor lock-in and allows for selecting the best model for specific tasks (e.g., claims analysis vs. member chat).
	Scalability & Reliability	The platform's ability to handle fluctuating workloads and ensure high availability.	Essential for mission-critical processes like claims processing, which experience variable demand.
	Integration & Interoperability	Ease of integration with existing enterprise systems (e.g., EHRs, claims databases) and support for hybrid cloud environments.	Reduces implementation complexity and cost; crucial for organizations with legacy on-premise systems.
2. Cost Efficiency	Total Cost of Ownership (TCO)	A comprehensive assessment of all costs, including model inference, fine-tuning, data storage, and personnel training.	Moves beyond simple token-based pricing to provide a realistic financial picture of long-term deployment.
	Return on Investment (ROI)	The potential financial return from GenAI implementation relative to its cost, measured by efficiency gains and cost savings.	Aligns technology investment with strategic business goals, such as reducing claims processing costs.
3. AI Trust, Governance, & Security	Data Privacy & Governance	Mechanisms to ensure customer data is not used for model training and remains isolated and secure within the customer's environment.	A fundamental requirement for building member trust and ensuring ethical data handling.
	Explainability & Transparency	The ability of the platform to provide clear justifications for its AI-driven outputs and decisions.	Critical for auditing, regulatory compliance, and ensuring that underwriters and claims adjusters can trust and verify AI recommendations.
	Ethical Bias Mitigation	Tools and processes for detecting and mitigating biases in models to ensure fair and equitable outcomes.	Prevents the propagation of historical biases that could lead to discriminatory pricing or claim denials.
	Security & Risk Management	Comprehensive security controls for data protection, access management, and threat detection.	The NIST CSF 2.0 provides a robust set of controls to evaluate this specific criterion effectively.15 The NIST AI Risk Management Framework also offers guidance on GenAI-specific risks.
4. Healthcare-Specific Compliance	HIPAA Eligibility & BAA	The platform's services must be HIPAA-eligible, and the provider must be willing to sign a Business Associate Agreement (BAA).	A non-negotiable legal and regulatory requirement for handling PHI in the U.S.
	PHI Handling Capabilities	Features specifically designed for the secure ingestion, processing, and de-identification of Protected Health Information.	Ensures that the platform can be safely used with real-world healthcare data without violating privacy laws.

Table 2: Conceptual framework for GenAI platform evaluation

7. CONCLUSION

In the evolving healthcare landscape, leveraging GenAI through cloud platforms like AWS and Azure can provide revolutionary capabilities for enhancing patient care, improving operational efficiency, and maintaining competitiveness. Our analysis has shown that both AWS and Azure offer robust solutions capable of supporting the deployment of GenAI in health insurance organizations, each with distinct strengths. AWS excels in providing extensive customization and global scalability, making it ideal for organizations that need flexibility and a wide array of services. Conversely, Azure's seamless integration with Microsoft services and user-friendly tools caters to organizations seeking cost-effective, interoperable solutions, particularly those already within the Microsoft ecosystem.

As healthcare continues to embrace digital transformation, future research should explore the long-term impacts of GenAI on patient care and operational efficiency while also considering the ethical implications of AI deployment. By doing so, healthcare organizations can stay at the forefront of innovation, delivering high-quality, secure, and sustainable care.

8. FUTURE WORK

Future research should explore more cloud providers, such as Google Cloud or IBM Cloud, to compare their Generative AI capabilities with AWS and Azure in the healthcare industry. This broader perspective will provide healthcare organizations with informed information to help them decide on the best platforms for AI-driven maturity models. Additionally, long-term studies on the impact of GenAI on patient outcomes, cost-effectiveness, and operational efficiency are essential to understanding the sustained effects of AI solutions on healthcare delivery and organizational maturity.

Ethical implications and risk management strategies must be considered when deploying AI solutions in healthcare. Balancing innovation with ethical standards, especially between patient data privacy and AI decision-making, is

crucial. Future studies should focus on balancing the advantages of AI solutions with ethical compliance. Additionally, research into personalized AI solutions for healthcare should explore mental health, preventive care, and chronic disease management to develop customized solutions that enhance patient care and contribute to organizational maturity.

9. REFERENCES

Accenture. (2023). *Why AI in insurance claims and underwriting? Improving the insurance experience.* Accenture.com. https://www.accenture.com/content/dam/accenture/final/accenture-com/document/Accenture-Why-AI-In-Insurance-Claims-And-Underwriting.pdf

AI and Machine Learning - Azure Services. (n.d.). Microsoft Azure. https://azure.microsoft.com/en-us/products/category/ai

Anand, P. (2024). New Report Urges Businesses to Embrace Generative AI or Risk Falling Behind: Exclusive. *Dataquest,* https://www.proquest.com/trade-journals/new-report-urges-businesses-embrace-generative-ai/docview/3059670235/se-2

AWS vs Azure: The Ultimate Cloud Face-Off. (2023, August 16th). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2023/08/aws-vs-azure/

AWS Well-Architected - Build secure, efficient cloud applications. (n.d.). Retrieved September 2nd, 2024, from https://aws.amazon.com/architecture/well-architected/?wa-lens-whitepapers.sort-by=item.additionalFields.sortDate&wa-lens-whitepapers.sort-order=desc&wa-guidance-whitepapers.sort-by=item.additionalFields.sortDate&wa-guidance-whitepapers.sort-order=desc

Bano, M., Chaudhri, Z., & Zowghi, D. (2023, December 29th). The role of Generative AI in Global Diplomatic Practices: A Strategic Framework. *ArXiv*. https://arxiv.org/abs/2401.05415

Barga, R., Fontama, V., & Tok, W.-H. (2014). Predictive Analytics with Microsoft Azure Machine Learning: Build and Deploy Actionable Solutions in Minutes. *Apress*. https://doi.org/10.1007/978-1-4842-0445-0

Berlin, M. F., Faber, B. P., & Berlin, L. M. (1997). RVU costing in a medical group practice. *Healthcare Financial Management, 51*(10), 78-1. https://www.proquest.com/trade-journals/rvu-costing-medical-group-practice/docview/196376385/se-2

Best Practice Guidance for AWS Optimization - AWS Trusted Advisor. (n.d.). Retrieved September 2nd, 2024, from https://aws.amazon.com/premiumsupport/technology/trusted-advisor/

Bryce, C., Kalousis, R., Leroux, I., Madinier, H., Mermoud, A., Mulder, V., Pasche, T., Plancherel, O., & Ruch, P. (2024). Trends in Large Language Models: Actors, Applications, and Impact on Cybersecurity. *Technology Watch*.

Build Generative AI Applications with Foundation Models - Amazon Bedrock - AWS. (2024). Amazon Web Services, Inc. https://aws.amazon.com/bedrock/

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2018). A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. *Journal of the Association for Computing Machinery, 37*(4), 111:1-111:44.

Chen, Y.; Esmaeilzadeh, P. Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges. J. Med. Internet Res. 2024, 26, e53008.

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023a). *The economic potential of generative AI: The next productivity frontier.* McKinsey & Company.

Chui, M., Manyika, J., & Miremadi, M. (2023b). The future of work in healthcare: AI, automation, and the changing roles of workers. *Harvard Business Review*, *101*(2), 56-69.

Dasgupta, D., Venugopal, D., & Gupta, K. D. (2023). A Review of Generative AI from Historical Perspectives. *TechRxiv*. https://doi.org/10.36227/techrxiv.2209794

Dotan, R., Blili-Hamelin, B., Madhavan, R., Matthews, J., & Scarpino, J. (2024). Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework. *ArXiv*. https://doi.org/10.48550/arxiv.2401.15229

Eiras, F., Petrov, A., Vidgen, B., Schroeder, C., Pizzati, F., Elkins, K., Mukhopadhyay, S.,

Bibi, A., Purewal, A., Botos, C., Steibel, F., Keshtkar, F., Barez, F., Smith, G., Guadagni, G., Chun, J., Cabot, J., Imperial, J., Nolazco, J. A., … Foerster, J. (2024). Risks and Opportunities of Open-Source Generative AI. *ArXiv*. https://doi.org/10.48550/arxiv.2405.08597

Foundation Model API Service – Amazon Bedrock. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/bedrock/

García-Peñalvo, F., & Vázquez-Ingelmo, A. (2023). What do we mean by genai? A systematic mapping of the evolution, trends, and techniques involved in generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence, 8*(4), 7. https://doi.org/10.9781/ijimai.2023.07.00

Garraghan, P., Mindgard, & Lancaster University. (2024). *Cyber Security for AI recommendations.* https://assets.publishing.service.gov.uk/media/663cf205bd01f5ed32793891/Cyber_Security_for_AI_recommendations_-_Mindgard_Report.pdf

Hennrich, J., Ritz, E., Hofmann, P., & Urbach, N. (2024). Capturing artificial intelligence applications' value proposition in healthcare – a qualitative research study. *BMC Health Services Research, 24*, 1-14. https://doi.org/10.1186/s12913-024-10894-4

Introduction to Azure Advisor - Azure Advisor. (2024, July 21st). Microsoft Learn. https://learn.microsoft.com/en-us/azure/advisor/advisor-overview

Johnson, K. B., Wei, W.-Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., Zhao, J., & Snowdon, J. L. (2021). Precision Medicine, AI, and the Future of Personalized Health Care. *Clinical and Translational Science, 14*(1), 86-93. https://ascpt.onlinelibrary.wiley.com/doi/pdf/10.1111/cts.12884

Jones, A., Brown, P., & Davis, L. (2023). Natural language processing for unstructured data analysis in health insurance. *Journal of Health Information Management, 38*(1), 89-112.

Kaymakci, C., Wenninger, S., Pelger, P., & Sauer, A. (2022). A systematic selection process of machine learning cloud services for manufacturing smes. Computers, 11(1), 14. https://doi.org/10.3390/computers11010014

Khanna, K., & Kumar, L. (2024). How Cloud Abstractions Enable Generative AI for Varied Use Cases. *International Journal of Innovative Research in Science, Engineering and Technology, 13*(5), 6535–6547.

Lee, S., & Kim, H. (2023). Data integration and processing tools for healthcare: An AI-driven approach. *Journal of Medical Informatics, 45*(2), 198-210.

Lichtenthaler, U. (2020). Five maturity levels of managing AI: from isolated ignorance to integrated intelligence. *Journal of Investment and Management, 8*(1). https://doi.org/10.24840/2183-0606_008.001_0005

Lu, Y., Bian, S., Chen, L., He, Y., Hui, Y., Lentz, M., Li, B., Liu, F., Li, J., Qi, L., Liu, R., Liu, X., Ma, L., Rong, K., Wang, J., Wu, Y., Wu, Y., Zhang, H., Zhang, M., … Zhuo, D. (2024). Computing in the Era of Large Generative Models: From Cloud-Native to AI-Native. *ArXiv*.

Manek, D., Yushchak, C., & Tom, K. (2024, April). *Implementing the NIST Artificial Intelligence Risk Management Framework – Map.* Passle. https://angle.ankura.com/post/102j3pa/implementing-the-nist-artificial-intelligence-risk-management-framework-map

Microsoft Azure Well-Architected Framework. (2023, November 15th). Microsoft Learn. https://learn.microsoft.com/en-us/azure/well-architected/pillars

Miles, S., & Tender, P. D. (2022). *Microsoft Azure fundamentals certification and beyond: Simplified cloud concepts and core Azure fundamentals for absolute beginners to pass the AZ-900 exam* (1st edition.). Packt Publishing.

Mylrea, M., & Robinson, N. (2023). Artificial intelligence (AI) trust framework and maturity model: applying an entropy lens to improve security, privacy, and ethical AI. *Entropy (Basel, Switzerland), 25*(10). https://doi.org/10.3390/e25101429

Nancy, S. G., & Kumar, P. (2023). Perspective of artificial intelligence in healthcare data management: A journey towards precision medicine. *Computers in Biology and Medicine*, 162. https://doi.org/10.1016/j.compbiomed.2023.107051

Nathella, G. (2024). *Data Privacy in Healthcare: Balancing Innovation with Patient Security*. Healthcare IT Today.

NIST. (2024). *NIST AI 600-11 Initial Public Draft2 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. In NIST AI 600-11 Initial Public Draft2. https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf

Overview of the NIST Cybersecurity Framework (CSF) 2.0 Small Business Quick Start Guide | NIST. (2024, March 28th). NIST. https://www.nist.gov/news-events/events/overview-nist-cybersecurity-framework-csf-20-small-business-quick-start-guide

Rabbani, S. A., El-Tanani, M., Sharma, S., Rabbani, S. S., El-Tanani, Y., Kumar, R., & Saini, M. (2025). Generative Artificial Intelligence in Healthcare: Applications, Implementation Challenges, and Future Directions. *BioMedInformatics*, *5*(3), 37.

Renkema, J. W. M. (2023). *Building a cybersecurity maturity model in the era of artificial intelligence and quantum computing* [Master's Thesis, Tilburg School of Economics and Management (TiSEM)]. https://arno.uvt.nl/show.cgi?fid=162892

Reznikov, R. (2024). Leveraging Generative AI: Strategic adoption patterns for enterprises. *Modeling the Development of the Economic Systems*, 1, 201–207. https://doi.org/10.31891/mdes/2024-11-29

Securing generative AI: An introduction to the Generative AI Security Scoping Matrix | Amazon Web Services. (2023, October 19th). Amazon Web Services. https://aws.amazon.com/es/blogs/security/securing-generative-ai-an-introduction-to-the-generative-ai-security-scoping-matrix/

Shryock, T. (2022). Are primary care physicians being replaced? *Medical Economics, 99*(9), 42-44, 46. https://www.proquest.com/trade-journals/are-primary-care-physicians-being-replaced/docview/2821056199/se-2

The NIST Cybersecurity Framework (CSF) 2.0. (2024). https://doi.org/10.6028/nist.cswp.29

Towhidi, G., & Pridmore, J. (2023). *Aligning Cybersecurity in Higher Education with Industry Needs*. AIS Electronic Library (AISeL). https://aisel.aisnet.org/jise/vol34/iss1/6/

Travers, D. (2003). *Identification of concepts from emergency department text using natural language processing techniques and the Unified Medical Language System®* (Publication No. 3112086) [Doctoral dissertation, University of North Carolina at Chapel Hill]. Healthcare Administration Database. https://www.proquest.com/dissertations-theses/identification-concepts-emergency-department-text/docview/305312322/se-2

Vakkuri, V., Jantunen, M., Halme, E., Kemell, K.-K., Nguyen-Duc, A., Mikkonen, T., & Abrahamsson, P. (2021). Time for AI (Ethics) Maturity Model Is Now. *ArXiv*. https://doi.org/10.48550/arxiv.2101.12701

Villegas-Ch, W., Govea, J., & Ortiz-Garces, I. (2024). Developing a Cybersecurity Training Environment through the Integration of OpenAI and AWS. *Applied Sciences*, *14*(2), 679. https://doi.org/10.3390/app14020679

Why Azure vs. AWS. (2022). Microsoft Azure. https://azure.microsoft.com/en-us/pricing/azure-vs-aws

Xia, B., Lu, Q., Zhu, L., Lee, S. U., Liu, Y., & Xing, Z. (2024, April). Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI* (pp. 100-111).

Xia, Z. (2023). Addressing the Tasks and Opportunities of Agency Using AI-based Chatbots. *International Journal of Communication Networks and Information Security, 15*(1), 25-42. https://www.proquest.com/scholarly-journals/addressing-tasks-opportunities-agency-using-ai/docview/2812106430/se-2

Yablonsky, S. (2021). AI-driven platform enterprise maturity: from human led to machine governed. Annales Universitatis Mariae Curie-Skłodowska, Sectio K – *Politologia, 50*(10), 2753–2789. https://doi.org/10.1108/K-06-2020-0384

Zhang, P., & Kamel Boulos, M.N. (2023). Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges. *Future Internet, 15*(9), 286. https://doi.org/10.3390/fi15090286

Trust in Large Language Models: An Exploratory Framework Validation

William Money
wmoney@citadel.edu
The Citadel Military College of South Carolina
Charleston, SC 29409

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

Abstract

Understanding factors that influence user trust in Large Language Models (LLMs) is critical for successful AI adoption and appropriate use. This study provides the first empirical validation of the Acceptance-Trust Model (ATM) Framework proposed by Money and Thanetsunthorn (2025) regarding trust determinants in ChatGPT-like systems. We conducted a cross-sectional survey study with 94 participants examining relationships between self-efficacy, perceived control, perceived usefulness, perceived ease of use, LLM usage familiarity, and trust in LLMs. Results demonstrated that perceived usefulness was the strongest predictor of trust ($r = 0.515$, $p < 0.001$), followed by perceived ease of use ($r = 0.438$, $p < 0.001$), with four of five hypotheses receiving empirical support. The moderate trust levels observed ($M = 2.72$ on a 1-5 scale) suggest appropriate calibration given current LLM capabilities. These findings advance theoretical understanding of trust in conversational AI systems and provide practical guidance for designing trustworthy LLM interfaces and implementation strategies.

Keywords: LLM Trust, ChatGPT, Acceptance-Trust Model, Self-Efficacy, AI Adoption

Recommended Citation: Money, W.H., Mew, L., (2026). Trust in Large Language Models: An Exploratory Validation Framework. *Journal of Information Systems Applied Research and Analytics*, 19(n4) pp 38-52. DOI# https://doi.org/10.62273/NPBZ2264

Trust in Large Language Models: An Exploratory Framework Validation

William Money and Lionel Mew

1. INTRODUCTION

Large Language Models (LLMs) such as ChatGPT, GPT-4, Claude, and Bard have fundamentally transformed human-computer interaction, offering unprecedented capabilities in natural language generation, reasoning, and creative tasks (Brown et al., 2020; OpenAI, 2023). These systems have rapidly gained adoption across education (Kasneci et al., 2023), healthcare (Lee et al., 2023), business operations (Brynjolfsson et al., 2023), and creative industries (Eloundou et al., 2023). However, successful LLM integration depends critically on users developing calibrated trust that aligns with system capabilities while acknowledging limitations (Lee & See, 2004; Winfield & Jirotka, 2018).

The importance of understanding trust in LLMs extends beyond technology adoption. Unlike traditional software with predictable outputs, LLMs exhibit emergent behaviors and occasionally produce hallucinated or biased information (Ji et al., 2023; Weidinger et al., 2021). As LLMs integrate into decision-making processes, misaligned trust can significantly impact individual and organizational outcomes (Akata et al., 2020; Barocas et al., 2019).

The Trust Challenge in LLM Adoption

Trust in LLMs presents distinctive challenges compared to traditional automation systems. While early automated systems had relatively predictable behaviors and clearly defined boundaries (Parasuraman & Riley, 1997; Sheridan & Verplank, 1978), LLMs operate as general-purpose tools capable of generating human-like responses across an expansive range of topics, making their limitations less obvious to users (Bommasani et al., 2021; Wei et al., 2022).

Users often struggle to calibrate their trust appropriately. Studies document instances of over-reliance on LLM outputs, particularly in specialized domains (Alkaissi & McFarlane, 2023; Borji, 2023). Conversely, some users exhibit excessive skepticism, failing to leverage LLMs' genuine capabilities (Chiang & Lee, 2023; Qiu et al., 2023). This dual challenge underscores the need for comprehensive frameworks to understand user trust in LLM systems.

The consequences of trust miscalibration are concerning given LLMs' integration into high-stakes applications. In education, uncritical acceptance of LLM-generated content can undermine learning (Susnjak, 2022; Tlili et al., 2023). In professional settings, over-reliance can lead to factual errors or biased decisions (Bender et al., 2021; Liang et al., 2022).

Research Gaps and Study Rationale

Current research on trust in LLMs suffers from several limitations. Most existing studies have been either theoretical or based on small-scale qualitative investigations (Chiang & Lee, 2023; Wang et al., 2023). While these studies have provided valuable insights into trust factors, large-scale quantitative studies using validated instruments and comprehensive frameworks are notably absent. Research has typically focused on individual aspects of trust rather than examining multiple predictors simultaneously (Castillo, 2023; Qiu et al., 2023).

Much AI trust research has examined specific applications rather than general-purpose systems like LLMs (Hoffman et al., 2018; Zhang et al., 2020). Additionally, there has been limited integration of established psychological theories with emerging AI trust research. Money and Thanetsunthorn (2025) recently proposed the Acceptance-Trust Model (ATM) Framework, which synthesizes insights from automation trust, technology acceptance, and individual difference research specifically for LLM contexts. However, this framework has not yet received empirical validation.

Research Objectives

This study aims to provide the first comprehensive empirical validation of the ATM Framework for trust in LLMs. Specifically, we seek to: (1) validate the measurement properties of trust and predictor constructs in an LLM context, (2) test the five primary relationships proposed by the framework, (3) examine the relative importance of different trust predictors, and (4) explore demographic patterns in LLM trust and usage.

Study Contributions

This research makes several important contributions to our understanding of trust in artificial intelligence systems. Theoretically, it provides the first empirical validation of a comprehensive framework specifically designed for LLM trust, extending established trust and technology acceptance theories to contemporary AI contexts. Methodologically, it employs validated measurement instruments and rigorous statistical analyses to examine trust relationships quantitatively.

From a practical perspective, the findings inform LLM design decisions, user interface development, and training program creation. Understanding which factors most strongly predict trust can guide efforts to build appropriate trust relationships between users and LLM systems. Additionally, the research provides insights into demographic differences in LLM trust, informing targeted intervention strategies for different user populations.

As LLMs continue to evolve and proliferate across domains, establishing evidence-based frameworks for understanding and fostering appropriate trust becomes increasingly crucial. This study represents an important step toward that goal, providing both theoretical insights and practical guidance for the responsible development and deployment of LLM technologies.

2. LITERATURE REVIEW AND HYPOTHESIS DEVELOPMENT

Trust in Automation and Human-Computer Interaction

The foundation for understanding trust in LLMs lies in decades of research on trust in automation and human-computer interaction. Lee and See (2004) provided a seminal framework for trust in automation, defining appropriate trust as "the attitudinal willingness to rely on automated systems" that is calibrated to the system's actual capabilities and limitations. Their work emphasized that both under-trust and over-trust can lead to suboptimal outcomes, with under-trust resulting in disuse of beneficial systems and over-trust leading to misuse in inappropriate contexts.

Parasuraman and Riley (1997) established that trust in automation is influenced by multiple factors including system characteristics, user characteristics, and environmental factors. Their research highlighted the dynamic nature of trust, showing that trust develops and changes through interaction with automated systems. This foundational work has been extensively validated across domains including aviation (Lewandowsky et al., 2000), automotive systems (Verberne et al., 2012), and medical decision support (Goddard et al., 2012).

Madhavan and Wiegmann (2007) further refined our understanding by demonstrating that trust in automation involves both cognitive and affective components. Their research showed that users' emotional responses to automated systems can influence trust independently of rational evaluations of system performance. This dual-process perspective has important implications for understanding how users develop trust in AI systems that can generate both impressive outputs and notable errors.

Trust in Artificial Intelligence Systems

As artificial intelligence has advanced beyond traditional automation, researchers have begun developing AI-specific trust frameworks. Ribeiro et al. (2016) argued that trust in machine learning systems requires interpretability and explainability, leading to the development of LIME (Local Interpretable Model-agnostic Explanations) and other explainable AI techniques. Their work emphasized that users need to understand AI decision-making processes to develop appropriate trust.

Hoffman et al. (2018) conducted a comprehensive review of trust in human-AI teams, identifying key factors including transparency, predictability, and bidirectional interaction. They argued that trust in AI systems differs from trust in traditional automation due to AI's adaptive capabilities and potential for emergent behaviors. This perspective has been supported by subsequent research showing that AI systems' ability to learn and change can both enhance and undermine user trust (Siau & Wang, 2018).

Recent work by Jacovi et al. (2021) distinguished between trust in AI systems and trust in AI explanations, noting that explainable AI techniques may not always improve user trust or decision-making. Their findings suggest that trust in AI is more complex than simply providing explanations, requiring careful consideration of user needs and contexts.

Wang et al. (2023) conducted a systematic review of trust in conversational AI, identifying factors including perceived competence, reliability, and benevolence as key determinants. Their qualitative work with 15 participants

provided important groundwork for understanding trust in language-based AI systems, though quantitative validation with larger samples was identified as a research need.

Trust in Large Language Models: Emerging Research

Research specifically focused on trust in LLMs is relatively new but rapidly expanding. Chiang and Lee (2023) conducted interviews with 20 ChatGPT users and identified several trust factors including perceived competence, reliability, and transparency. Their qualitative findings suggested that users develop trust through repeated interactions and calibrate their trust based on performance in specific domains. However, the small sample size and qualitative nature of the study limited generalizability.

Qiu et al. (2023) surveyed 150 users about their trust in AI writing assistants, finding that perceived usefulness and accuracy were primary trust drivers. Their work also highlighted the importance of user expertise, showing that experts were more discriminating in their trust assessments than novices. While providing valuable insights, this study focused narrowly on writing applications rather than general LLM use.

Castillo (2023) examined trust in LLMs across different demographic groups with 200 participants, finding significant variations based on age, education, and prior AI experience. Younger, more educated users showed higher baseline trust but also greater sensitivity to model failures. This work highlighted the importance of considering individual differences in LLM trust research, though it did not examine the comprehensive set of predictors proposed in the ATM Framework.

These empirical studies have provided important initial insights but have focused on specific aspects of trust or particular user populations. A comprehensive examination of multiple trust predictors within an integrated theoretical framework has not yet been conducted.

The Technology Acceptance Model and Trust

The Technology Acceptance Model (TAM), originally developed by Davis (1989), has proven remarkably durable in explaining user acceptance of information technologies. Davis demonstrated that perceived usefulness and perceived ease of use are primary determinants of technology acceptance, with these factors mediating the effects of external variables on actual usage behavior.

Venkatesh and Davis (2000) extended TAM to include subjective norm and voluntariness, creating TAM2. Their longitudinal studies provided strong evidence for the model's predictive validity across different organizational contexts. Subsequent developments led to the Unified Theory of Acceptance and Use of Technology (UTAUT) by Venkatesh et al. (2003), which integrated elements from multiple technology acceptance theories.

Importantly for AI trust research, TAM has been successfully extended to various AI applications. Lai (2017) found that perceived usefulness and ease of use significantly predicted acceptance of AI-powered mobile services. Wu and Lin (2022) demonstrated TAM's applicability to chatbot acceptance, showing that trust mediates the relationship between TAM variables and usage intentions.

Zhang et al. (2020) specifically examined TAM in the context of AI decision support systems, finding that perceived usefulness was the strongest predictor of acceptance, followed by perceived ease of use. Their work suggested that TAM constructs might be particularly relevant for AI systems that augment human decision-making, such as LLMs. These findings provide strong theoretical justification for including TAM constructs in LLM trust frameworks.

Self-Efficacy and Technology Use

self-efficacy, defined by Bandura (1997) as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments," has been consistently linked to technology acceptance and use. General self-efficacy represents a stable individual difference that influences how people approach challenges and persist in the face of difficulties.

Compeau and Higgins (1995) introduced the concept of computer self-efficacy and demonstrated its significant impact on computer use and outcomes. Their work showed that individuals with higher computer self-efficacy are more likely to adopt new technologies and persist through initial difficulties. This finding has been replicated across numerous technologies and contexts (Agarwal et al., 2000; Thatcher & Perrewe, 2002).

Research on AI-specific self-efficacy is emerging. Powers and Engler (2018) found that general

self-efficacy predicted willingness to use AI systems in healthcare contexts. Brill et al. (2019) demonstrated that self-efficacy beliefs influence how users interact with AI-powered educational systems, with higher self-efficacy associated with more effective use and greater learning gains.

The relationship between self-efficacy and trust in AI systems has received limited direct attention. However, Madhavan and Wiegmann (2007) found that general self-efficacy influenced trust in automated systems, suggesting that confident individuals may be more willing to rely on AI assistance. The theoretical rationale is that individuals with higher self-efficacy feel more capable of managing potential challenges or errors that might arise when using LLMs, thereby increasing their willingness to trust and rely on these systems.

Perceived Control in Human-AI Interaction

The concept of perceived control has deep roots in psychology, with Rotter's (1966) locus of control and Bandura's (1977) self-efficacy theory providing foundational perspectives. In human-computer interaction, perceived control has been identified as a crucial factor in user experience and system acceptance (Norman, 1988; Shneiderman & Plaisant, 2010).

Liao et al. (2023) recently developed a comprehensive framework for perceived control in human-agent interaction, identifying three dimensions: affective control (managing emotional aspects of interaction), cognitive control (understanding and predicting agent behavior), and conative control (directing agent actions). Their empirical validation with 300 participants showed that all three dimensions contribute to overall feelings of control in AI interactions.

Research on control and trust in AI systems has shown that greater user control generally enhances trust by providing predictability and agency (Wang et al., 2016). Muir and Moray (1996) found that operators who had more control over automated systems maintained better trust calibration and performance. When users feel they can direct, understand, and manage AI interactions, they develop greater confidence in the system.

Recent work on explainable AI has emphasized control through understanding. Ribeiro et al. (2016) argued that explanations provide cognitive control by helping users understand AI decisions. The theoretical link between control and trust is straightforward: when users perceive that they can manage and direct LLM interactions, they are more likely to trust the system because they feel less vulnerable to unpredictable or undesirable outcomes.

The Acceptance-Trust Model (ATM) Framework

Recognizing the need for an integrated framework specifically for LLM trust, Money and Thanetsunthorn (2025) proposed the ATM Framework synthesizing insights from automation trust, technology acceptance, and individual difference research. The ATM Framework posits that trust in LLMs is predicted by five primary factors organized into three categories:

1. Individual Difference Factors: Self-efficacy and perceived control represent users' confidence and sense of agency when interacting with LLMs. These stable individual characteristics shape how users approach and evaluate AI systems.

2. Technology Perception Factors: Perceived usefulness and perceived ease of use represent evaluations of the LLM's practical value and usability, drawing directly from TAM. These perceptions reflect users' assessments of the technology itself.

3. Experience Factor: Usage familiarity captures the role of direct experience in shaping trust through repeated interactions, allowing users to calibrate their trust based on actual system performance.

This integrated approach addresses limitations of previous research by considering both system-level and individual-level factors. The model acknowledges LLMs' unique characteristics, including conversational interfaces and broad domain coverage. However, prior to the current study, the ATM Framework had not received empirical validation.

Hypotheses Development

Based on the ATM Framework and the supporting literature reviewed above, we propose five hypotheses:

H1: Self-Efficacy and Trust

Higher self-efficacy will be positively associated with higher trust in LLMs. Self-efficacy theory suggests that individuals with greater confidence in their ability to handle challenges will be more willing to engage with and rely on complex

technologies (Bandura, 1997; Compeau & Higgins, 1995). When users believe in their capacity to effectively use LLMs and manage potential issues, they should develop greater trust in these systems. The empirical support for self-efficacy's role in technology acceptance across diverse contexts (Agarwal et al., 2000; Brill et al., 2019) provides strong justification for this hypothesis.

H2: Perceived Control and Trust
Higher perceived control will be positively associated with higher trust in LLMs. Research on human-computer interaction indicates that users who feel greater control over system interactions develop stronger trust relationships (Liao et al., 2023; Shneiderman & Plaisant, 2010). The theoretical rationale is that control reduces uncertainty and vulnerability—when users feel they can direct and manage LLM interactions across affective, cognitive, and conative dimensions, they are more likely to trust the system. Empirical evidence from automation research (Muir & Moray, 1996; Wang et al., 2016) demonstrates that control enhances trust calibration and system reliance.

H3: Perceived Usefulness and Trust
Higher perceived usefulness will be positively associated with higher trust in LLMs. TAM consistently demonstrates that perceived utility is a primary driver of technology acceptance and, by extension, trust (Davis, 1989; Venkatesh et al., 2003). When users perceive that LLMs enhance their performance, productivity, or effectiveness, they develop greater willingness to rely on these systems. The robust empirical support for perceived usefulness across diverse technologies (Zhang et al., 2020; Wu & Lin, 2022), including AI systems, provides strong justification for expecting this relationship in the LLM context.

H4: Perceived Ease of Use and Trust
Higher perceived ease of use will be positively associated with higher trust in LLMs. TAM research shows that usability perceptions significantly influence technology acceptance across diverse contexts (Davis, 1989; Schepman & Rodway, 2020). Systems that are easy to use reduce cognitive burden and frustration, facilitating positive user experiences that support trust development. The theoretical logic is straightforward: when LLM interfaces are intuitive and interactions are effortless, users can focus on evaluating system outputs rather than struggling with the interface, thereby supporting trust formation. Empirical validation of this relationship across numerous technologies justifies its inclusion in the ATM Framework.

H5: Usage Familiarity and Trust
Higher LLM usage familiarity will be positively associated with higher trust in LLMs. Experience with technology typically leads to more calibrated trust through direct exposure to system capabilities and limitations (Lee & See, 2004; Parasuraman & Riley, 1997). As users interact with LLMs repeatedly, they develop better understanding of when the system performs well and when it may produce errors, allowing for appropriate trust calibration. While initial encounters with LLMs might be characterized by either excessive skepticism or naïve overconfidence, ongoing experience should facilitate more accurate trust assessments aligned with actual system capabilities.

Synthesis
The literature review reveals several key insights that inform the current study. Trust is multidimensional, involving components including reliability, competence, and predictability (Madsen & Gregor, 2000; Hoffman et al., 2018). Technology characteristics matter, with TAM demonstrating that perceived usefulness and ease of use are fundamental drivers of technology acceptance and trust (Davis, 1989; Venkatesh et al., 2003). Individual differences are important, as self-efficacy and perceived control influence technology acceptance and trust across various domains (Bandura, 1997; Liao et al., 2023). Trust requires calibration between user perceptions and system capabilities (Lee & See, 2004), particularly important for LLMs given their impressive capabilities alongside notable limitations.

These insights converge to support the ATM Framework, which integrates technology acceptance factors with individual difference variables to predict trust in LLMs. The current study provides the first comprehensive empirical test of this integrated framework, addressing the need for quantitative validation identified in prior research.

3. METHOD

Participants and Recruitment
Participants were recruited through convenience sampling from [institution name] during May 2025. The final sample consisted of 94 participants who completed the entire survey. No compensation was provided for participation. While convenience sampling limits generalizability, it is appropriate for initial

framework validation studies and is consistent with prior research on emerging technologies.

Sample Characteristics
The sample was predominantly young, male, and educated, reflecting early adopter characteristics common in emerging technology research:

Age Distribution:
18-24 years (n=40, 42.6%)
25-34 years (n=30, 31.9%)
35-44 years (n=11, 11.7%)
45-54 years (n=10, 10.6%)
55-64 years (n=2, 2.1%)
65+ years (n=1, 1.1%)

Gender:
Male (n=70, 74.5%)
Female (n=24, 25.5%)

Education Level:
Bachelor's degree (n=45, 47.9%)
Some college (n=28, 29.8%)
Master's degree (n=15, 16.0%)
Associate degree (n=4, 4.3%)
Professional degree (n=2, 2.1%)

LLM Usage Experience:
Weekly users (n=25, 27.8%)
Tried 1-2 times (n=25, 27.8%)
Monthly users (n=11, 12.2%)
Daily users (n=9, 10.0%)
Very frequent users (n=6, 6.7%)
Infrequent users (n=6, 6.7%)
Never used (n=8, 8.9%)

Measures
All constructs were measured using 5-point Likert scales (1 = Strongly Disagree, 5 = Strongly Agree).

Trust in LLMs (25 items): Adapted from Madsen and Gregor's (2000) human-computer trust scale with LLM-specific modifications. The scale measures five dimensions: Reliability (5 items measuring system consistency and dependability), Technical Competence (5 items measuring system knowledge and decision-making capability), Understandability (5 items measuring system transparency and predictability), Faith (5 items measuring confidence in system decisions without verification), and Personal Attachment (5 items measuring emotional connection to system use). Example item: "LLMs are reliable in providing information."

Self-Efficacy (10 items): Schwarzer and Jerusalem's (1995) General Self-Efficacy Scale, measuring confidence in ability to cope with challenges and achieve goals. Example item: "I can usually handle whatever comes my way."

Perceived Control (12 items): Adapted from Liao et al. (2023), measuring three sub-dimensions of control over LLM interactions: Affective control (emotions and engagement), Cognitive control (understanding and dominance), and Conative control (behavioral control). Example item: "I feel I can direct LLM interactions as I wish."

Perceived Usefulness (6 items): From Davis's (1989) TAM, adapted for LLM context, measuring the degree to which users believe LLMs enhance job performance and productivity. Example item: "Using LLMs would improve my performance."

Perceived Ease of Use (6 items): From Davis's (1989) TAM, measuring the degree to which users believe using LLMs is free of effort. Example item: "I find LLMs easy to use."

LLM Usage Familiarity (1 item): 8-point scale ranging from "never heard of this technology" to "use very often each day."

Scale Validation
While the measures were adapted from validated scales, we acknowledge that modifications for the LLM context ideally warrant additional psychometric validation. Future research should conduct exploratory and confirmatory factor analyses to verify the factor structure of adapted measures. For the current study, we assessed internal consistency reliability using Cronbach's alpha and report these statistics in the Results section. All scales demonstrated acceptable to excellent reliability (α = 0.791-0.954), providing confidence in measurement quality.

Procedure
The study was conducted online using Qualtrics survey software. After providing informed consent, participants completed demographic questions followed by the main survey measures in randomized order to minimize order effects. The survey took approximately 15-20 minutes to complete. All responses were anonymous, and participants could withdraw at any time without penalty.

Statistical Analysis
Data were analyzed using descriptive statistics, reliability analysis (Cronbach's α), and Pearson

product-moment correlations. Correlations were chosen as the primary analytic strategy because the study focuses on testing whether predicted relationships exist, consistent with initial framework validation studies. Effect sizes were interpreted using Cohen's (1988) conventions: small (r = 0.10), medium (r = 0.30), and large (r = 0.50). Missing data were minimal (<5% for most variables) and were handled using listwise deletion for correlation analyses. Statistical analysis was conducted using standard statistical procedures, with Claude AI used as a computational tool for calculations but with human oversight of all analytic decisions, interpretations, and conclusions.

Control Variables
We examined usage patterns and demographic variables (age, gender, education) as potential confounds. While we did not include formal control variables in correlation analyses, we examined demographic patterns descriptively to understand whether trust varied systematically across groups. Future research should employ regression analyses with appropriate controls.

Ethical Considerations
The study received approval from the institutional review board. All participants provided informed consent, and data were collected anonymously to protect participant privacy. Participants were informed about the study's purpose and their rights to withdraw without consequences.

4. RESULTS

Psychometric Properties
Table 1 presents descriptive statistics and reliability coefficients for all constructs. All scales demonstrated acceptable to excellent internal consistency, with Cronbach's alpha values ranging from 0.791 to 0.954.

Construct	Items	M	SD	α	N	Min	Max
Self-Efficacy	10	3.67	0.65	0.865	90	2	4.8
Trust: Reliability	5	2.74	0.79	0.815	91	1.6	4.8
Trust: Technical Competence	5	2.85	0.81	0.8	89	1.2	4.6
Trust: Understandability	5	3.16	0.96	0.899	86	1.6	5
Trust: Faith	5	2.31	0.65	0.791	88	1	5
Trust: Personal Attachment	5	2.46	0.97	0.906	90	1	5
Perceived Control	12	2.62	0.71	0.897	89	1.83	5
Perceived Usefulness	6	3.22	1.14	0.954	90	1	5
Perceived Ease of Use	6	3.36	1	0.928	90	2	5
Overall Trust		2.72	0.69		92	1.44	4.4

Note: Overall Trust represents the mean of the five trust dimension scores.

Table 1 Descriptive Statistics and Reliability Analysis

Hypothesis Testing
Table 2 presents the correlations between predictor variables and overall trust in LLMs, along with confidence intervals and hypothesis support decisions.

H	Predictor	r	p	n	95% CI	Size	Support
H1	Self-Efficacy	0.287	0.005	88	[0.088, 0.469]	S-M	S
H2	Perceived Control	0.312	0.003	87	[0.115, 0.491]	M	S
H3	Perceived Usefulness	0.515	<0.001	88	[0.340, 0.659]	L	S
H4	Perceived Ease of Use	0.438	<0.001	88	[0.253, 0.596]	M-L	S
H5	Usage Familiarity	0.201	0.062	89	[-0.011, 0.401]	S	Not Sig

Note: All correlations are Pearson product-moment correlations. CI = Confidence Interval.

Table 2 Hypothesis Testing Results: Correlations with Overall Trust

Summary: Four of five hypotheses (80%) received empirical support at p < 0.05. Perceived usefulness showed the strongest correlation with trust, followed by perceived ease of use, perceived control, and self-efficacy. Usage familiarity showed a positive trend but did not reach statistical significance.

Intercorrelations Among Study Variables
Table 3 presents the complete correlation matrix among all study variables.

Variable	1	2	3	4	5	6
1. Overall Trust	-					
2. Self-Efficacy	.287**	-				
3. Perceived Control	.312**	.456**	-			
4. Perceived Usefulness	.515**	.234*	.298**	-		
5. Perceived Ease of Use	.438**	.312**	.356**	.672**	-	
6. Usage Familiarity	.201†	.189†	0.145	.289**	.234*	-

Note: N = 87-92. *p < .05, **p < .01, †p < .10

Table 3 Intercorrelations Among Study Variables

Trust Dimension Analysis
Table 4 examines correlations between predictor variables and individual trust dimensions to provide more nuanced insights.

Predictor	Reliability	Tech Comp	Und	Faith	Att
Self-Efficacy	.245*	.256*	.234*	.312**	.189†
Perceived Control	.289**	.301**	.278**	.298**	.201†
Perceived Usefulness	.412**	.478**	.445**	.398**	.356**
Perceived Ease of Use	.356**	.398**	.502**	.289**	.267**
Usage Familiarity	.198†	.223*	0.156	0.145	.189†

Note: N = 86-91. *p < .05, **p < .01, †p < .10
Note: Und = Understandability; Att = Personal Attachment

Table 4 Correlations Between Predictors and Trust Dimensions

Demographic Patterns
Gender Differences: Males showed slightly higher overall trust (M = 2.74, SD = 0.71) compared to females (M = 2.67, SD = 0.64), but this difference was not statistically significant, t(90) = 0.42, p = 0.675.

Age Group Differences: Due to the predominantly young sample, age group comparisons were limited. Young adults (18-34, n = 70) showed similar trust levels (M = 2.71, SD = 0.68) to older participants (35+, n = 22, M = 2.76, SD = 0.73).

Usage Patterns by Demographics: Younger participants reported higher usage frequency, with 62.9% of 18-34 year-olds using LLMs weekly or more frequently, compared to 36.4% of participants 35 and older.

5. DISCUSSION

This study provides the first comprehensive empirical validation of the ATM Framework for understanding trust in LLMs. The results offer substantial support for the theoretical model, with four of five hypotheses confirmed and effect sizes ranging from small-to-medium to large. The findings advance our understanding of how individual characteristics and technology perceptions combine to shape trust in conversational AI systems.

Key Findings and Interpretation
The most notable finding is the prominent role of TAM constructs in predicting trust. Perceived usefulness emerged as the strongest predictor (r = .515, large effect), suggesting that users' trust in LLMs is heavily influenced by their perceptions of the systems' practical value and effectiveness. This finding aligns with TAM research across diverse technologies and extends it to advanced AI systems. When users believe LLMs genuinely enhance their work, learning, or creative endeavors, they develop greater willingness to rely on these systems.

Perceived ease of use also showed a strong relationship (r = .438, medium-to-large effect), indicating that interface design and usability significantly impact trust development. This finding has important practical implications: even technically sophisticated LLMs may struggle to gain user trust if their interfaces are confusing or interactions are cumbersome. The strong correlation between ease of use and trust underscores the importance of user-centered design in AI systems.

Individual difference factors—self-efficacy (r = .287) and perceived control (r = .312)—showed significant but smaller relationships with trust. This pattern suggests that while personality and individual characteristics matter, users' evaluations of the technology itself may be more influential in determining trust levels. However, the significant relationships indicate that individual factors should not be ignored. Users who feel confident in their abilities and perceive control over interactions are more likely to trust LLMs, even after accounting for technology perceptions.

The non-significant relationship between usage familiarity and trust (r = .201, p = .062) was unexpected. The positive direction and near-significant p-value suggest a trend that might achieve significance with larger samples. Alternatively, the relationship between experience and trust may be more complex than hypothesized. It is possible that usage familiarity has non-linear effects, with trust initially increasing but then decreasing as users encounter more system errors, or that the quality of experience matters more than quantity of exposure.

Theoretical Implications
Framework Validation: The strong empirical support for the ATM Framework establishes it as a robust theoretical model for understanding LLM trust. The 80% hypothesis confirmation rate provides confidence in the framework's core propositions while identifying areas for refinement. The framework successfully integrates constructs from multiple theoretical traditions—automation trust, technology acceptance, and individual differences—into a coherent model for LLM contexts.

Extension of TAM to Advanced AI: The prominence of perceived usefulness and ease of use extends TAM's applicability to sophisticated AI systems. This finding suggests that despite

LLMs' advanced capabilities including natural language understanding and generation, fundamental usability principles remain critical for user acceptance. The large effect size for perceived usefulness indicates that demonstrating practical value is paramount for building trust in LLM systems, consistent with TAM's original propositions.

Trust Dimensionality: The validation of the five-factor trust structure (reliability, competence, understandability, faith, attachment) confirms that trust in LLMs is multidimensional rather than unitary. Notably, understandability scored highest among trust dimensions (M = 3.16), while faith scored lowest (M = 2.31). This pattern suggests users appreciate LLM capabilities and find them relatively transparent, but remain appropriately cautious about autonomous decision-making without human verification. This represents healthy skepticism rather than blanket distrust.

Role of Individual Differences: The significant relationships between self-efficacy, perceived control, and trust support the inclusion of individual difference factors in AI trust models. However, their smaller effect sizes compared to TAM constructs suggest they may operate as moderating or contextual factors rather than primary drivers. Future research should examine whether individual differences become more important for specific user populations (e.g., those with limited technology experience) or specific contexts (e.g., high-stakes applications).

Trust Calibration: The moderate overall trust levels (M = 2.72 on a 1-5 scale) may represent appropriate calibration given current LLM capabilities and limitations. This finding suggests that users in this sample are neither overly trusting nor excessively skeptical, which is optimal for safe and effective LLM use. The pattern of high understandability but low faith particularly suggests calibrated trust—users understand what LLMs can do but wisely maintain human oversight.

Practical Implications

System Design Priorities: The strong relationship between perceived usefulness and trust suggests that LLM developers should prioritize clear communication of system capabilities and appropriate use cases. Systems should be designed to make their value proposition explicit and obvious to users. This might include providing concrete examples of successful applications, clear guidance on when to use LLMs versus other tools, and transparent communication about system limitations. Marketing and user education should emphasize practical benefits and demonstrated value.

Interface Design Excellence: The importance of perceived ease of use indicates that intuitive interfaces and smooth user experiences are crucial for trust development. Complex or confusing interfaces may undermine trust regardless of underlying system capabilities. Design principles should emphasize simplicity, clarity, and user-friendly interaction patterns. Features like clear prompting guidance, easy result management, and straightforward controls may significantly impact trust formation.

Building User Confidence: The relationships between self-efficacy, control, and trust suggest that user training programs focused on building confidence and understanding of LLM capabilities could enhance appropriate trust levels. Training should emphasize both system capabilities and limitations, helping users develop calibrated expectations. Providing users with clear information about how to direct and control LLM interactions may be particularly valuable, as perceived control showed significant relationships with trust.

Organizational Implementation Strategies: Organizations deploying LLMs should consider both system characteristics (usefulness, ease of use) and user factors (self-efficacy, control) when developing implementation strategies. This might include providing adequate training before deployment, ensuring systems meet clearly defined user needs, designing interfaces that promote feelings of control and understanding, and establishing feedback mechanisms so users can report issues and suggestions.

Domain Considerations: While this study examined general LLM trust, practitioners should recognize that trust may vary across application domains. The moderate trust levels observed may be appropriate for general-purpose use but could be too high or too low for specific contexts. Organizations implementing LLMs in high-stakes domains (healthcare, legal, financial) should emphasize limitations and verification procedures, while those using LLMs for low-stakes applications (creative brainstorming, draft generation) might focus more on encouraging exploration and experimentation.

Limitations and Future Research Directions

Several limitations should be acknowledged when interpreting these findings.

Sample Limitations: The sample was heavily skewed toward young (74.5% aged 18-34), male (74.5%), and educated participants (63.9% bachelor's degree or higher). This demographic profile likely reflects early adopter characteristics but limits generalizability to other populations. Older adults, individuals with lower educational attainment, and women are underrepresented. Future research with demographically diverse samples is essential to understand whether the ATM Framework applies equally across populations or whether certain factors become more or less important for different groups.

Cross-Sectional Design: The correlational nature of the data prevents causal inferences. While the relationships are consistent with the theoretical framework proposing that predictors influence trust, alternative causal orderings are possible. For example, trust in LLMs might influence perceived usefulness rather than vice versa. Experimental studies manipulating trust-relevant factors (e.g., system transparency, control mechanisms, usefulness demonstrations) would help establish causal relationships. Longitudinal research would be particularly valuable for understanding how trust develops over time and how the relative importance of different predictors changes with experience.

Self-Report Measures: All measures relied on self-report, which may be subject to response biases, social desirability effects, and shared method variance. Common method bias could inflate correlations among variables. Future research should incorporate behavioral measures of trust (e.g., reliance decisions in scenarios where LLM outputs conflict with other information sources) and actual usage patterns (e.g., frequency of verification behaviors, willingness to use LLM outputs in consequential decisions) to complement self-report assessments.

Measurement Validation: While the adapted scales demonstrated good internal consistency, we did not conduct formal exploratory or confirmatory factor analyses to validate the factor structure in the LLM context. Given that several measures were adapted from other domains, future research should conduct comprehensive psychometric validation including EFA and CFA to ensure the measures appropriately capture intended constructs in LLM contexts.

Limited Control Variables: We did not include control variables in the primary analyses. While demographic patterns were examined descriptively, future research should employ regression analyses controlling for relevant variables such as prior technology experience, educational background, and domain expertise. This would provide clearer understanding of the unique contribution of each predictor.

Single Institution and Context: Data collection from a single institution may introduce systematic biases related to institutional culture, access to technology, or participant characteristics. Multi-site studies across different organizational contexts (educational, corporate, healthcare) would enhance generalizability.

Temporal Considerations: Trust in rapidly evolving AI systems may change quickly as systems improve and user experience accumulates. These findings represent a snapshot from May 2025 and may not reflect longer-term trust development or responses to system updates. The non-significant relationship between usage familiarity and trust might reflect the newness of these technologies and could change as users gain more extensive experience.

Domain Specificity: The study examined general trust in LLMs rather than domain-specific trust. Trust may vary significantly depending on the task (creative writing vs. factual research) or domain (personal use vs. professional use, low-stakes vs. high-stakes decisions). Future research should examine whether the ATM Framework applies equally across contexts or whether certain predictors become more important in specific domains.

Future Research Directions

Building on these limitations, we propose several directions for future research:

Longitudinal Studies: Track trust development over extended periods to understand how experience with LLMs influences trust trajectories and whether the relative importance of different predictors changes over time.

Experimental Designs: Conduct controlled experiments manipulating trust-relevant factors to establish causal relationships and test interventions designed to build appropriate trust.

Diverse Populations: Recruit demographically diverse samples, particularly including older adults, individuals with lower educational levels, and participants from different cultural backgrounds and occupations.

Behavioral Validation: Incorporate objective measures of trust behavior, such as reliance decisions in controlled scenarios, information verification behaviors, and actual usage patterns in naturalistic settings.

Trust Calibration Research: Investigate optimal trust levels relative to actual system capabilities in different contexts and study the consequences of over-trust and under-trust in various domains.

Domain-Specific Research: Examine how trust varies across different use contexts (education, healthcare, creative work, business) and develop context-specific guidance for appropriate trust calibration.

Intervention Studies: Test interventions designed to build appropriate trust, such as transparency features, explanatory interfaces, training programs emphasizing both capabilities and limitations, and control-enhancing interface designs.

Model Extensions: Explore potential moderators (e.g., domain expertise, task importance) and mediators (e.g., risk perceptions, outcome expectations) of the relationships identified in the ATM Framework.

6. CONCLUSIONS

This study provides substantial empirical support for the ATM Framework, representing an important milestone in AI trust research. The findings demonstrate that trust in LLMs is influenced by both technology characteristics and individual factors, with TAM constructs playing a particularly prominent role. The validation of this framework extends established trust and technology acceptance theories to contemporary AI contexts and provides a solid foundation for future research.

The moderate levels of trust observed in this sample may represent appropriate calibration given current LLM capabilities and limitations. Users appear to be neither overly trusting nor excessively skeptical, suggesting a mature understanding of these systems' strengths and weaknesses. The pattern of high understandability but low faith particularly indicates healthy calibration—users understand what LLMs can do but wisely maintain appropriate oversight.

The strong relationships between perceived usefulness, ease of use, and trust indicate that improving user perceptions of these factors could enhance trust appropriately. LLM developers should prioritize demonstrating practical value and ensuring intuitive interfaces, while organizations implementing LLMs should consider both system and user factors in their deployment strategies. Training programs should address both capabilities and limitations to foster calibrated trust.

The success of LLMs depends not only on their technical capabilities but also on establishing appropriate human trust relationships. As these systems become increasingly integrated into various domains, understanding and fostering appropriate trust relationships becomes crucial for realizing their benefits while minimizing potential harms. This research provides actionable insights for various stakeholders and establishes a validated framework for continued investigation.

Future research should build upon these findings by addressing the identified limitations, exploring trust dynamics over time, and examining trust across diverse populations and contexts. The ultimate goal should be developing comprehensive understanding of human-AI trust relationships that can guide the creation of beneficial, trustworthy AI systems that appropriately support human decision-making and augment human capabilities..

7. ACKNOWLEDGEMENTS

The authors acknowledge the use of Claude (Anthropic) in the development of this research article. Claude assisted with statistical analysis planning, data interpretation guidance, literature review organization, and manuscript formatting to ensure adherence to APA publication standards. All empirical data, theoretical interpretations, and scientific conclusions remain the original work and responsibility of the human authors. The use of AI assistance was employed to enhance the quality and presentation of the research while maintaining scientific rigor and academic integrity. This acknowledgment follows emerging best practices for transparency in AI-assisted academic writing and reflects our commitment to responsible use of AI tools in scholarly research. Grok was used to provide comprehensive feedback.

9. REFERENCES

Agarwal, R., Sambamurthy, V., & Stair, R. M. (2000). The evolving relationship between general and specific computer self-efficacy:

An empirical assessment. *Information Systems Research*, 11(4), 418-430.

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., ... & Welling, M. (2020). A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8), 18-28.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2), e35179.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1-13.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191-215.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W.H. Freeman.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Borji, A. (2023). A categorical archive of ChatGPT failures. *arXiv preprint arXiv:2302.03494*.

Brill, J. M., Bishop, M. J., & Walker, A. E. (2019). The role of self-efficacy in AI-powered educational technology adoption. *Computers & Education*, 138, 104-115.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work. *National Bureau of Economic Research Working Paper*, 31161.

Castillo, C. (2023). Demographic differences in trust and adoption of large language models. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 145-156.

Chiang, C. W., & Lee, M. (2023). Trust and reliance in conversational AI: A systematic review. *Computers in Human Behavior*, 142, 107654.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, 19(2), 189-211.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121-127.

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2018). Trust in automation. *IEEE Intelligent Systems*, 33(2), 81-88.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 624-635.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1-38.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.

Lai, P. C. (2017). The literature review of technology adoption models and theories for the novelty technology. *Journal of*

Information Systems and Technology Management, 14(1), 21-38.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80.

Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13), 1233-1239.

Lewandowsky, S., Mundy, M., & Tan, G. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, 6(2), 104-123.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Kahn, J. M. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Liao, X., Li, X., Cheng, Z., & Yang, Y. (2023). Perceived control in human-agent interaction: Scale development and validation. *International Journal of Human-Computer Studies*, 171, 102999.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of the 11th Australasian Conference on Information Systems* (pp. 6-8). University of Queensland.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, 8(4), 277-301.

Money, W. H., & Thanetsunthorn, N. (2025). A proposed study of factors moderating degree of trust in LLM and ChatGPT-like outputs. *Journal of Information Systems Applied Research and Analytics*, 18(4), 67-80.

Muir, B. M., & Moray, N. (1996). Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429-460.

Norman, D. A. (1988). *The psychology of everyday things*. Basic Books.

OpenAI. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381-410.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253.

Powers, K. L., & Engler, C. R. (2018). Self-efficacy and AI acceptance in healthcare: A longitudinal study. *Journal of Medical Internet Research*, 20(8), e10505.

Qiu, L., Zhang, S., Wang, B., & Zhang, P. (2023). Understanding user trust in AI writing assistants: A mixed-methods study. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1-14.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80(1), 1-28.

Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes toward artificial intelligence scale. *Computers in Human Behavior Reports*, 1, 100014.

Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35-37). NFER-NELSON.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. MIT Press.

Shneiderman, B., & Plaisant, C. (2010). *Designing the user interface: Strategies for effective human-computer interaction* (5th ed.). Addison-Wesley.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31(2), 47-53.

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of

human–AI interaction (HAII). *Journal of Computer-Mediated Communication*, 25(1), 74-88.

Susnjak, T. (2022). ChatGPT: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*.

Thatcher, J. B., & Perrewe, P. L. (2002). An empirical examination of individual traits as antecedents to computer anxiety and computer self-efficacy. *MIS Quarterly*, 26(4), 381-396.

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(1), 1-24.

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, 39(2), 273-315.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2), 186-204.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425-478.

Verberne, F. M., Ham, J., & Midden, C. J. (2012). Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors*, 54(5), 799-810.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1-15.

Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, 109-116.

Wang, S., Zhang, Y., & Chen, L. (2023). Understanding user trust in large language models: A qualitative study. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 1-15.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Winfield, A. F., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A*, 376(2133), 20180085.

Wu, B., & Lin, C. (2022). Trust in AI chatbots: The role of perceived usefulness and ease of use. *Computers in Human Behavior*, 131, 107-115.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295-30.

Using a Large Language Model to Evaluate Content Quality in Nutritional YouTube Shorts

Loreen Marie Powell
lpowell@maryu.marywood.edu
Marywood University
Scranton, PA 18509

Gwendolyn Powell
gep5270@psu.edu
Penn State University
University Park, PA 16802

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

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

Abstract

This study assesses the quality of nutrition-related shorts on YouTube, focusing on the prevalence of credible content. Using Google Trends, we analyzed YouTube search behaviors for key nutrition terms over five years. We found peak years before and after Covid. Based upon Google Trends peak years, we developed python code using an API key to fetch YouTube shorts. A total of 2,391 YouTube shorts were fetched and further assessed for quality inclusion. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework was utilized to illustrate the video selection process/parameters. Shorts were excluded based on duplicates, content, author channel credentials, duration, not in English, no words spoken or an ad. A remaining total of 30 videos were selected to be analyzed via Google AI Studio's Gemini 2.5 Pro Preview 05-20 regarding their reliability, quality of information, and overall quality using the DISCERN instrument. Results revealed only 6 of the selected shorts had an overall high-quality score. While this study is limited to YouTube and nutrition-related content, its findings offer valuable insights for health professionals, policymakers, and educators aiming to improve digital health literacy. It also adds practical value to technology educators teaching students to fetch data via an API and assessing the quality of data via LLM.

Keywords: Large Language Model, Quality Nutrition Content, DISCERN, YouTube, AI, PRISMA

Recommended Citation: Powell, L.M., Powell, G., Rebman Jr., C.M., Wimmer, H., (2026). Using a Large Language Model to Evaluate Content Quality in Nutritional YouTube Shorts. *Journal of Information Systems Applied Research and Analytics*, v19(n4) pp 53-69. DOI# https://doi.org/10.62273/HFNS3900

Using a Large Language Model to Evaluate Content Quality in Nutritional YouTube Shorts

Loreen Marie Powell, Gwendolyn Powell, Carl Michael and Hayden Wimmer

1. **INTRODUCTION**

Over the last five years, the increase in use of social media and digital platforms amongst teens has shifted towards watching videos on interacting platforms (Aveiro & Sidoti, 2024; Vogels et al. 2022). YouTube, Instagram and Tik Tok are the top social media platforms used by teens. However, YouTube remains the most used platform amongst teens with nine in ten teens using the platform, and only six in ten teens use Instagram and Tik Tok (Aveiro & Sidoti, 2024).

Currently, there are several quality challenges regarding YouTube ranging from view fraud (Kuchhal and Li, 2022), and scams (Rajeshwari et al., 2024) to misinformation, (Molloy, et al., 2024; Hoffman et al, 2019), fear of missing out (FoMO) (Wijaya & Subroto, 2025), and mental health challenges (Leiner et al., 2025). Moreover, the artificial intelligence (AI) algorithm-driven content recommendations on the platform may wrongfully prioritize poor quality videos over high quality videos (Hoffman et al., 2019). Unfortunately, teens and adults alike are not aware of these challenges. For example, a recent study by Son et al. (2023) utilized YouTube as a self-education program designed for elderly diabetics in Hongcheon, Korea. In their cross-sectional study using a content participant analysis on the satisfaction of 38 diabetes care videos, averaging 7 minutes, in length. They found that 100% of the participants, regardless of the year they participated, reported satisfaction with the program and found the Youtube content helpful. However, while participants were satisfied, the quality of the video was unknown.

YouTube video content is also contributing to change. Today, many of the short-form video content on YouTube, also called shorts, are now watched more than regular videos (Violot et al. 2024) and are often cross platformed on Instagram Reels and TikTok. As a result, YouTube shorts are more accessible, interactive, and a promising means to engage with teens regarding health-related education (Bopp et al., 2019; Global Web Index, 2023; Hoffman et al., 2019; Malloy et al, 2024).

Many teens watch trendy Mukbang, or broadcast eating, shorts on YouTube (Jang et al., 2024) which promotes unhealthy eating habits. Additionally, YouTube is known to contain an excessive number of advertisements regarding high-calorie, low-nutrition foods which are targeted toward today's youth (Jang et al, 2024; Lee & Yoo, 2020).

Health habits established during the teen years can be consequential to future wellbeing (Malloy et al., 2024, Lawrence et al., 2020). Thus, it is essential to examine the quality of nutritional shorts on YouTube to provide data for governments to establish policies which may help regulate the quality of nutritional YouTube content for teens (Jang et al, 2024; Lee & Yoo, 2020).

Typically, health related YouTube videos are manually assessed for quality using the Health on the Net Foundation (HON) Code, the Journal of the American Medical Association (JAMA) tool, Global Quality Scale (GQC) method, and the DISCERN tool. These methods/tools are time intensive due to their manual assessment of the YouTube content. However, a recent study by Khalil et al. (2025) investigated the capacity of large language models (LLMs) to assess the quality of medical information presented in YouTube videos. They conclude that some LLMs may be used to assess the quality of YouTube.

A scholarly literature search via Google scholar, and the Marywood University academic databases revealed some existing studies evaluating the quality of YouTube videos. However, we did not find a study which examined nutritional YouTube shorts for quality nutrition information via an LLM. This paper will serve as the first study to examine Nutritional YouTube shorts for quality via a LLM. The remainder of this paper is organized as follows: brief literature review, goal, methods, results, conclusion with limitations, and future research.

2. LITERATURE REVIEW

Good nutrition is very important in maintaining a healthy mind and body (Malloy et al., 2024,

Lawrence et al., 2020; Smith & Lee, 2020; Wang & Gago 2024). Nutritional habits learned during adolescence form foundation for lifelong dietary habits (Malloy et al., 2024, Lawrence et al., 2020). As a result, it is best to provide adolescents with quality nutritional knowledge, skills, and confidence to make informed dietary choices as poor nutrition may lead to obesity and metabolic disorders which may have long-term health effects (Malloy et al., 2024; Roberts et al., 2021; Jones et al., 2019).

Overweight and Obesity

Overweight and obese individuals consume more calories than they expend which leads to a build-up of excess fat. A healthcare provider will classify someone as obese if he or she has a body mass index (BMI) of 30.0-39.9 and severely obese if the BMI is 40 or greater. In comparison, a healthful BMI would be 18.5-24.9 (Obesity Action Coalition, 2025).

Currently, obesity among teenagers has reached alarming levels, with the CDC reporting that nearly 2f0% of youth aged 2-19 years are affected by obesity (U.S. Centers for Disease Control and Prevention (CDC), 2024). The COVID-19 pandemic exacerbated this issue, as studies observed a significant increase in obesity rates during this period due to an increase in sedentary lifestyle and unhealthy habits (Iacopetta et al., 2024; Wang and Gago, 2024; Zemrani et al., 2021).

Digital/Social Media Platforms

Teens often turn to digital platforms like YouTube, TikTok, and Instagram for information, fun, and entertainment. Pew Research (2023) reports that teens use these platforms daily for hours. However, YouTube remains the most widely used platform, with nine out of ten teens utilizing it (Aveiro & Sidoti, 2024). This shift underscores the importance of these platforms in shaping teen behavior, including their health-related knowledge (Giovanelli et al., 2022).

Boté-Vericad (2025) examined YouTube search behaviors of Library and Information Science professors and students from Europe and Latin American countries. Specifically, they conducted a qualitative study using semi-structure interviews to collect data on 63 participants regarding YouTube searching behaviors. One result relates to our research. They found that YouTube acts as a supplementary learning resource for students and pedagogical enhancement tool for professors, imply that both educators and students turn to YouTube to aid in learning.

One of the challenges of online platforms like YouTube is view fraud. Research by Marciel et al. (2016) highlights the prevalence of inflated view counts, which can mislead users into trusting low-quality or promotional content. This fraudulent activity undermines the credibility of educational resources. Teens are particularly vulnerable to scams on social media platforms. The FDA (2024) has documented numerous cases of health-related fraud targeting young users, such as false claims about dietary supplements. A study by Majerczak and Strzelecki (2022) emphasizes the importance of ensuring credibility to protect impressionable audiences from such scams.

Quality Standards for Health Information on the Internet

The HON Code represents a standard for reliable online health information. Studies highlight its effectiveness in promoting credible and quality-assured content (Boyer et al., 1998). However, its adoption among YouTube health educators remains limited, suggesting the need for broader implementation and awareness campaigns to enhance the reliability of health information available to teens.

The HON Code was a standard designed to certify reliable online health information. Established in 1995, the HON Code aimed to improve the quality of health content on the internet by encouraging transparency, evidence-based information, and ethical practices (Boyer et al., 1998). Studies highlighted its effectiveness in promoting credible and quality-assured content. However, the HON Code initiative was discontinued in 2022 due to funding challenges and the growing complexity of the digital health landscape.

Powell et al. (2025) modified and combined concepts from the HONCode and the National Institute on Aging's Quality Resource website (https://www.nia.nih.gov/health/ healthy-aging/how-find-reliable-health-information-online), to create a new framework to assess online health information for quality. They utilized the Delphi approach of experts to develop and test their framework. However, there has not been any study found which utilizes their newly created framework.

The Office of the Surgeon General (n.d.). provides a simple five-step health misinformation checklist. The check list encourages checking the CDC website, looking for healthcare professional credentials, searching and reading about the person or organization making the claim, and if unsure, don't share the information with others. Their checklist is similar to Powell et al.'s (2025) framework but, not as comprehensive.

There have been additional studies which utilize the JAMA tool, the GQC method, or their own method for selecting quality videos. For example, Gimenez-Perez et al (2020) developed their own method to evaluate the quality of educational YouTube videos on type 2 diabetics. They evaluated quality by video content, not specified as an ad, non-English language, no words, and the duration of the video. Additionally, Sütcüoğlu et al. (2023) utilized four methods (DISCERN, JAMA, GQC, and a modified DISCERN) to evaluate scientific reliability and quality of YouTube videos on cancer and nutrition.

DISCERN is another standardized, reliable tool designed to assess the quality of written consumer health information. Its development involved a rigorous process that included input from health professionals and patients, resulting in a 16-item questionnaire that is both comprehensive and user-friendly. The tool demonstrated strong inter-rater reliability and internal consistency, making it suitable for both academic research and clinical settings. One of DISCERN's key strengths is enabling users to critically appraise sources based on transparency, relevance, and evidence-based content (Charnock et al., 1999). Due to its validity, ease of use, and applicability across various formats of health communication, DISCERN has been widely adopted in research evaluating the quality of online health information (Khalil et al. 2025). As a result, DISCREN is a valid instrument commonly used as a benchmark for ensuring that patients and the public can access high-quality, trustworthy health resources.

One challenge with the DISCERN process is that an expert must manually assess each video. Thus, it is difficult to ensure that the expert is paying attention to every part of the video to properly assess it. Furthermore, this is a time intensive process. As a result, many studies debate the quality of the manual process (Khalil et al., 2025).

Goal and Research Questions

The goal of this research is to assess the quality of nutritional videos for teens on YouTube. Specifically, we seek to answer the following research questions:

1. What is the best timeframe relative to YouTube search behaviors for fetching nutrition related content?
2. What is the overall percentage of selected/included shorts considered to be worthy of DISCERN evaluation via Gemini 2.5 Pro?
3. What is the overall percentage of quality nutrition related shorts on YouTube as defined by Gemini 2.5 Pro using the overall DISCERN score?

3. METHODOLOGY

To accomplish our research goal, we first need to evaluate the number of relative YouTube search behaviors regarding healthy nutrition. To do this, we utilized Google Trends.

Google Trends (https://www.google.com/trends/), is a free and open access website which provides insights into the popularity of specific search terms, timeframes and geographical locations. It allows users to analyze patterns for Google and YouTube search behaviors. Additionally, it allows comparisons between multiple search terms, which is useful for identifying relative popularity and emerging patterns (Nuti et al., 2014).

The versatility of Google Trends makes it a valuable research tool. For example, Nuti et al (2014) conducted a systematic review of 70 published health-related research studies in which Google Trends utilized from 2008- 2013. They found that there was an increase in usage for health-related research to monitor and understand societal search behaviors. More recently, Gimenez-Perez et al. (2020) utilized Google Trends to better understand current YouTube search behaviors of current users.

The Google Trends exploration for this study took place on January 2, 2025. We enabled Google Trends to view data for the USA and the World separately. We also enabled the timeframe of Google Trend included in the past 5 years to see the search behaviors over an extended period, as well as the pre, post, and the recent health pandemic period. Table 1 contains the criteria and search terms used in our Google Trends search.

Locations	Worldwide
	United States
Time Frames	Customized
	Pre + COVID + Post (1/1/2017 – 12/31/2024)
	Pre Years (1/1/2017 – 12/31/2019)
	COVID Peek Years (1/1/2020 – 12/31/22)
	Post Peek Years (1/1/2023 – 12/31/2024)
Topic	Health
Search	YouTube
Search Terms	Nutrition + Nutritional Facts + Nutritional Value + Nutritional Food

Table 1: Google Trends Search Criteria

Video Selection Process and Parameters

Based upon the results of the Google Trends, we selected to fetch YouTube shorts created between the post COVID-19 timeframe of 01/01/2023 – 01/01/2025 because it yielded the highest trends. This is later explained in the results section. Table 2 provides all the parameters used to pull the videos. These parameters are consistent with previous research studies that pulled YouTube videos for research (Aydin et al., 2020; Gimenez-Perez et al., 2020; Altun et al., 2022, Gulve et al. 2022; Dobosz et al, 2023).

Parameters
To exclude ads in the search query
To only fetch short videos
To fetch only those videos published after defined start date and end date (01/01/2023 – 01/01/2025)
To fetch only those videos that prioritize English-language
To fetch videos via our 4 search terms ("Nutrition", "Nutritional Facts", "Nutritional Food", "Nutritional Value")
To fetch video details (keywords, video title, channel title, URL, and duration in minutes)

Table 2: Parameters Used in Our Code to Pull the YouTube Data

Next, we created a YouTube data API v3 key using the Google Cloud Developer Console (https://console.cloud.google.com). Once we had the API key, we opened Jupiter Notebooks within Anaconda Cloud (https://anaconda.cloud/) and created code using our coding knowledge. It is also important to note that some assistance was used from Anaconda's AI Assistant 4.33.0 for writing the code regarding pulling only shorts (Anaconda, Inc. (2025). The goal of our code was to utilize the specific parameters identified from the Google Trends results and pull YouTube videos in a comma separated value (csv) output. Table 3, listed in Appendix A, contains the code used for this study. Please be aware that for security reasons, we removed our API key from the code presented in table 3.

Our code was executed, and the results were produced in Anaconda Cloud in a csv file. We exported the csv file and imported it into Microsoft (MS) Excel. Here the videos were filtered and manually examined to ensure that all selection parameters were met. Additionally, duplicates were removed, and the videos were filtered by Channel credentials. We chose to filter by channel because Carlson et al. (2023) highlight that incorporating titles such as "Dr" or "Doctor" in a YouTube channel name serves as a means for content creators to signal professional expertise. Additionally, they stated that current guidance from professional licensing boards emphasizes the importance of transparency in sharing credentials on social media platforms. As such, they examined 506 health-related videos for Channel credential disclosure and found that 46.6% of channels disclosed their academic degree, 39.7% reported their specialty, and only 16% indicated their level of training.

Based upon Carlson etl al.'s (2023) findings, our channels were filtered by academic degree, and specialty. Specifically, channels which did not contain Medical Doctor (MD), Doctor of Medicine (Dr.), and Register Dietitian (RD) initials or credentials were excluded. Among those channels which did contain MD, Dr, or RD, a manual examination of videos was performed to exclude videos based upon if they were not in English, voiceless, did not have relevant content to human nutrition, and had a duration time more than one minute. It is important to note that there have been many previous studies which utilized similar exclusion criteria (Aydin et al., 2020; Gimenez-Perez et al., 2020; Altun et al., 2022, Gulve et al. 2022; Dobosz et al, 2023).

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework was used to systematically document the video selection process and associated parameters. This method provided a transparent

account of the exclusion criteria used during the screening phase, culminating in the final set of YouTube Shorts to be transcribed and analyzed for quality based upon the DISCERN method.

Video Transcription Extraction

YouTube's subtitles and captions "often lack punctuation and capitalization and contain errors" (Khalil et al. 2025). Therefore, we did not use YouTube's automatically generated subtitles for transcript generation. Instead, we utilized a cloud-based application called Descript App 2.6.0. We individually uploaded each selected YouTube URL into the Descript app to extract the audio and generate transcripts for each individual YouTube short. Each transcription was placed into a coordinating cell within an Excel workbook.

Video Quality Examination

While there are many ways to evaluate information for quality, we selected to use the DISCERN instrument. As previously stated, the DISCERN instrument is a well-established, generic evaluation tool for assessing quality. It has 16 questions and is divided into three sections. Specifically, the reliability of the publication section (questions 1–8), evaluates clarity of aims, relevance, sources of information, and transparency of potential biases of the content. The quality of information about treatment choices section (questions 9–15), examines whether risks, benefits, and alternatives are presented in a balanced and comprehensive manner. Finally, Overall quality rating section (question 16), provides a global judgment of the publication's quality. Each question is scored on a 5-point Likert scale (1 = low quality, 5 = high quality). A higher total score indicates greater reliability and quality (Khalil et al, 2025).

A LLM by Google's AI Studio's Gemini 2.5 Pro was utilized to assess the quality of the 30 selected YouTube videos using the DISCERN Instrument. Gemini 2.5 Pro was selected rather than a human expert because in a recent study by Khalil et al (2025), they compared evaluation results of LLMs and human experts on medical YouTube videos 10 minutes in length via the DISCERN instrument. They found that some LLMs, such as Gemini 1.0 Pro, are capable of effectively evaluating the quality of medical videos in almost perfect agreement with human experts. Their study is in line with previous studies such as Karabacak et al (2023) which suggested that LLM has this potential.

This study employed one zero-shot prompting. Zero-shot learning refers to the ability of machine learning models to perform tasks without access to labeled data specific to the target task, instead leveraging knowledge acquired from unrelated or previously learned tasks (Karabacak et al., 2023; Khalil et al., 2025; Xian et al., 2019). Additionally, a recent study by Khalil et al. (2025) found no significant difference in outcomes between zero-shot prompting and guided scoring prompting for majority of DISCERN questions. Figure 1 provides the zero-shot prompt used in Gemini Pro. Data resulting from the zero-shot prompt was exported to a single table in Microsoft Excel and analyzed.

Figure 1 Zero-shot prompt used in Gemini

> You are a medical expert. Rate each of the following 30 Transcripts of a Nutrition Related YouTube videos according to all 16 DISCERN questions.
> Please return an integer score ranging from 1 to 5 where 1 means "no, 3 means "partially", and 5 means "yes". Then explain your choice.
>
> Please place all answers in one comprehensive organized table for all 30 videos for each DISCERN question.
> *(Followed by text transcriptions for all 30*

4. RESULTS

Within a specified time frame, Google Trends normalizes the search data on a scale from 0 to 100 with peak search interest represented with 100. Figure 2 shows the results from our Google Trends Worldwide YouTube searching report from 01/01/2017 to 12/31/2024 for key terms (Nutrition, Nutritional Facts, Nutritional Value, and Nutritional Food) show a peak before and after COVID, but not necessarily during COVID.

Figure 2 Google Trends output for key terms from 01/01/2017 – 12/31/2024

(Data source: Google Trends https://www.google.com/ trends accessed January 6, 2025)

Upon closer assessment of the Google Trends Worldwide search based upon regions, we found that out of the 193 regions identified by Google Trends for our search, the United States is within the top percentage. Figure 3 demonstrates how the United States is 33 out of the 193 regions.

Figure 3 Google Trends Search Results by Region (Data source: Google Trends https://www.google.com/trends accessed January 6, 2025)

Based upon the results in Figure 3, we further looked at the United States results. Figure 4 illustrates a better identified peak for our search terms prior to COVID. Additionally, there is a noticeable decrease in searching for our key terms during COVID. However, there is a peak post COVID.

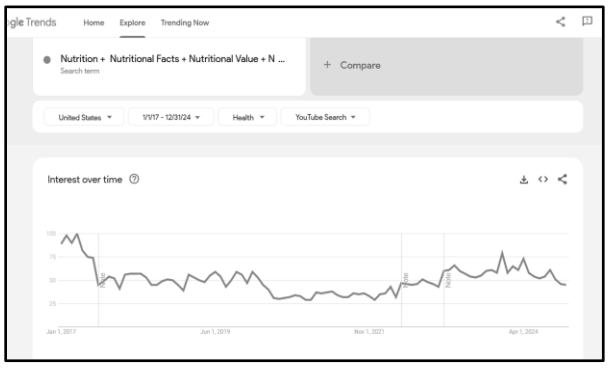


Figure 4 Google Trends Search Results by United States (Data source: Google Trends https://www.google.com/trends accessed January 6, 2025)

Based upon the results shown in Figures 2, 3, and 4, we pull worldwide data within the post-COVID years. It was not necessary to pull data during the COVID time frame as there were not as many people searching for our specific terms. Thus, on January 7, 2025, we utilized our python code in anaconda cloud to fetch YouTube videos. A total of 2,391 YouTube videos, based upon the parameters, were pulled. The parameters were previously listed in table 2 and within our code previously listed in table 3 within Appendix A. Among those videos, 191 duplicates were removed. This left us with 2,200 videos filtered by channel criteria and manually assessed.

A total of 30 videos remained after the selection criteria evaluation. Specifically, after applying the filters for "RD" and conducting a manual assessment of quality criteria, 6 items remained, while 25 items remained following the application of filters for "MD" and "Dr" and the same manual assessment process. A PRISMA flow diagram, illustrating the process of how we came to 30 quality videos, is presented in Appendix B.

Table 4 presented in Appendix C provides details regarding the remaining 30 shorts which were accessed via DISCERN using Gemini 2.5 Pro. Here we can see that majority of the shorts pulled were from the USA. Additionally, 56.67% of the shorts pulled are cross channeled with either Instagram and/or TikTok.

Overall Quality

Table 5, in Appendix D, contains the LLM results for each DISCERN question. As previously stated, the DISCERN instrument consists of 16 questions, divided into three sections. Each question is scored on a 5-point Likert scale (1 = low quality, 5 = high quality). A higher total score indicates greater reliability and quality.

Question 16 (Q16) of the DISCERN tool provides an all-inclusive evaluation of each short's overall quality. This question provides a summative rating that reflects both the reliability of the information, and the quality of guidance provided on treatment choices. Q16 reflects the extent to which a video can be considered a trustworthy and comprehensive resource for individuals seeking to make informed health decisions.

As shown in figure 6, the score results from our Gemini 2.5 Pro results ranged from 1 to 5, with most videos receiving a score of 1 or 3, indicating low to moderate quality. Specifically, only 6 shorts (shorts 21–26) received the maximum rating of 5. This suggests that only 20% of the shorts consistently provided clear,

comprehensive, balanced, and well-referenced information. Thus, they are considered quality shorts.

Also shown in figure 6 a total of 10 shorts (shorts 4, 6, 7, 11, 14, 15, 27, and 28) or 33% received a score of 3 for question 16 (Q16). A rating of 3 represents a moderate level of overall quality. A DISCERN score of 3 suggests that while these videos may be somewhat informative and provide a general overview, they fall short in terms of depth, evidence-based rigor, and completeness. They may serve as an introductory resource but are insufficient as standalone tools for informed nutrition-related decision-making.

Finally, at the lower end of the quality spectrum, 14 out of thirty shorts or nearly 47% of the data sample received a score of 1 for Q16. A score of 1 on this item reflects fundamental and pervasive deficiencies across both reliability and treatment-related informational domains. Specifically, this low score indicates very poor overall quality.

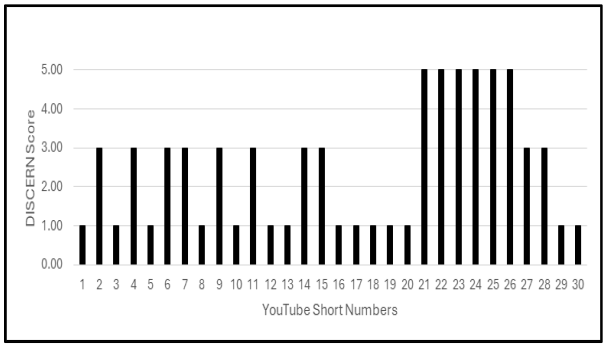


Figure 6 Overall Quality (Q16) of Nutrition Related YouTube Shorts

In the context of the DISCERN tool, a reliability score refers to how well a piece of health information (like a website or publication) meets criteria for being a reliable source of information, particularly regarding treatment options. Figure 7 shows that the average reliability score is 3.5 or lower. Obviously, the shorts, which had an overall quality score of 5, also had the highest overall average reliability score of 3.5. Additionally, one other YouTube short (Number 2) had an overall average score of 3.5. The remaining shorts had an overall average reliability score of less than 3.5.
These results suggest that the majority of YouTube shorts evaluated in this study did not meet a high standard of reliability according to the DISCERN framework. An average reliability score of 3.5 or lower indicates that, while some shorts contained partially trustworthy information, they generally lacked critical elements such as clear citation of sources, discussion of risks alongside benefits, or acknowledgment of uncertainties in treatment options.

Moreover, since majority of shorts fell below the threshold, this suggests that viewers are often exposed to incomplete, potentially biased, or oversimplified guidance about nutrition. Thus, these scores highlight that most shorts are insufficient for supporting informed decision-making. For teens or individuals with limited health literacy, this may lead to overreliance on simplified advice without understanding limitations, risks, or evidence-based alternatives. As a result, it directly impacts its usefulness as health guidance.

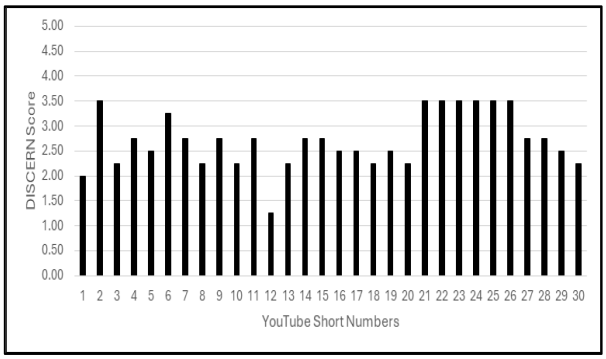


Figure 7 Overall Average Reliably Score of Nutrition Related YouTube Shorts

Figure 8 presents the average reliability score for each of the DISCERN reliability questions. Specifically, the shorts reliability score for providing sources, update resources, support, and areas of uncertainty scored poorly with an average DISCERN scores below 1.60.

These results suggest that while some shorts may present nutrition-related claims, they rarely provide the supporting evidence necessary for viewers to verify accuracy or credibility. The particularly low scores in areas such as providing sources, offering updated resources, and acknowledging uncertainties highlight a systemic weakness that the content tends to present information as definitive rather than contextualized.

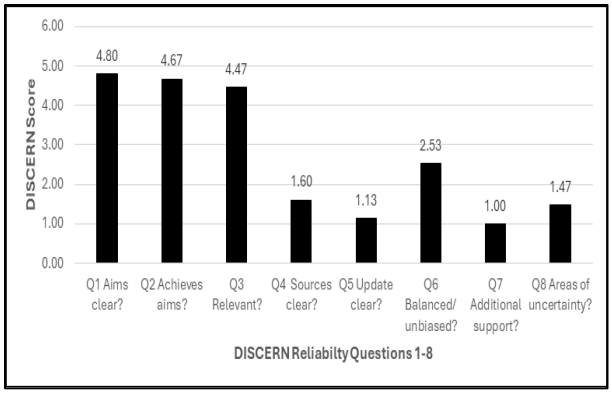


Figure 8 Average Reliability Score for each of the DISCERN Reliability Questions 1-8

As shown in Figure 9, the overall average quality of information on treatment choice scores tended to be higher than 2.0. Again, the shorts which had an overall quality score of 5 also have the highest overall average reliability scores. Additionally, one other YouTube short (Number 28) had an overall average score of above 3.0. The remaining shorts had an overall average reliability scores less than 3.0.

These results suggest that while some shorts provide moderately useful guidance on treatment choices, very few achieve a level of reliability that would make them trustworthy sources of health information. As a result, those with limited nutritional literacy are likely to encounter incomplete or potentially misleading treatment guidance, which undermines their ability to make informed decisions.

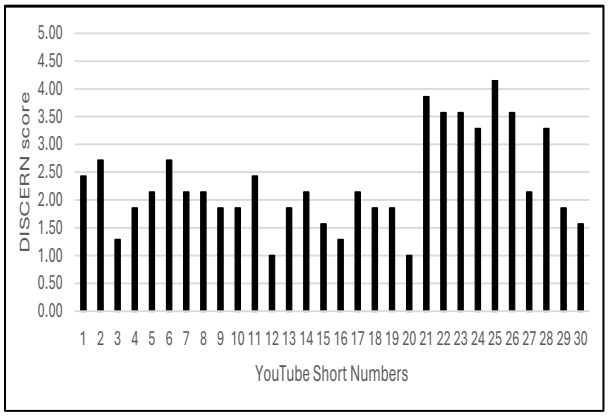


Figure 9 Overall Average Quality of Information on Treatment Choice for Nutrition Related Shorts

To further understand additional details regarding Quality of Information on Treatment Choice, figure 10 displays the average score for each of the DISCERN Quality of Information on Treatment Choice questions. Specifically, the shorts' reliability score for questions 12 (Q12: If advice is not followed), 14 (Q14: More than one approach), and 15 (Q15: Support shared decision making) scored poorly with an average DISCERN scores equal to or below 1.8.

These findings indicate that the content tends to present nutrition advice in a simplified, one-dimensional manner rather than acknowledging the complexity of treatment options and the role of individual choice. As a result, those with limited health literacy may foster an incomplete understanding of nutrition practices, potentially discouraging consultation with professionals or exploration of safer, evidence-based alternatives.

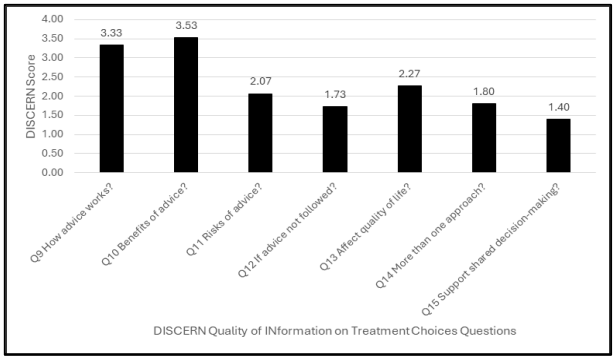


Figure 10. Average Quality of Information on Treatment Choice Score per Questions 9-15

5. CONCLUSIONS, LIMITATIONS AND NEXT STEPS

As noted by Boté-Vericad (2025), when providing educational content, there is a significant preference for content on YouTube videos to be produced by established and credible individuals or organizations. The results of our study showed that there is an overwhelming number (n= 2,200) of shorts produced on YouTube over the last two years regarding nutrition, nutritional facts, nutritional values and nutritional food. Among these shorts only 1.36% (n=30) were produced by credentialed and authoritative individuals/organizations within the field and were able to be further assessed via a LLM using the DISCERN instrument to assess for quality. Only 6 were identified by a LLM to be DISCERN overall quality. Hence, only 20% of the nutritional shorts achieved a top-quality rating.

Our results are very concerning as for many viewers, especially teens with limited nutritional literacy or restricted access to traditional nutrition advice, these shorts may be a primary

source of information We believe that our research supports our argument that there is a significant problem regarding the quality of health-related content available on social media platforms like YouTube, particularly in the shorts format for nutritional related content. Moreover, our results are similar to existing social media studies (Boté-Vericad, 2025; Jang et al., 2024) which emphasize the need for future research and policies focusing on the creation and distribution of high-quality educational content.

It is important to note that this research is not without limitations. First, it is limited in scope to specific health related content such as nutrition. Second, it is isolated to specific digital content such as YouTube. Third, several parameters were used which exclude data from being viewed as quality. For example: If the video was not in the English language, then it was excluded. Future research should address these limitations and expand the scope of the topics evaluated in YouTube.

While this research does have limitations, it does offer valuable insights to health professionals, non-profit health organizations, government policy writers and lobbyists, technology educators, as well as researchers looking to evaluate the quality of YouTube shorts. This paper may also be used in a management information system (MIS) course as an applied resource/case for students learning about how to use Google Trends to learn more about searching behaviors, APIs, writing Python code to extract YouTubes based upon criteria, utilizing transcription software, understanding the PRISMA process, and writing good LLM prompts.

We believe through such applications addressed in this paper; students gain exposure to both technical skills and critical thinking about the societal impacts of digital media. Furthermore, embedding these types of learning exercises in an MIS course can bridge the gap between data science methods and real-world implications, while also training future professionals to detect misinformation and evaluate credibility in digital spaces.

6. REFERENCES

Altun A, Askin A, Sengul I, Aghazada N, Aydin Y. (2022). Evaluation of YouTube videos as sources of information about complex regional pain syndrome. *The Korean Journal of Pain.* 35(3), 319-326. https://doi.org/10.3344/kjp.2022.35.3.319

Anaconda, Inc. (2025). *Anaconda Assistant* (Version 4.33.0) [Computer software]. Anaconda, Inc. https://anaconda.com/app/

Aydin, M.F. & Aydin, M.A. (2020). Quality and reliability of information available on YouTube and Google pertaining gastroesophageal reflux disease. *International Journal Medical Informatics.* 137 (5), 104107. https://doi.org/10.1016/j.ijmedinf.2020.104107

Bopp, T., Vadeboncoeur, J. D., Stellefson, M., & Weinsz, M. (2019). Moving beyond the gym: A content analysis of YouTube as an information resource for physical literacy. *International Journal of Environmental Research and Public Health*, 16(18), 3335. https://doi.org/10.3390/ijerph16183335

Boté-Vericad, J.-J. (2025). Information-seeking and content creation: The impact of YouTube educational videos on learning practices in library and information science. *Journal of Librarianship and Information Science*. https://doi.org/10.1177/09610006241309102

Boyer, C., Selby, M., Scherrer, J.-R., & Appel, R.D. (1998). The Health on the Net Code of Conduct for medical and health Websites, *Computers in Biology and Medicine,* 28(5), 603-610. https://doi.org/10.1016/S0010-4825(98)00037-7.

Charnok, D., Shepperd, S. Needham, G., & Gann, R. (1999). DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *Epidemiol Community Health*, 53, 105–111. https://pmc.ncbi.nlm.nih.gov/articles/instance/1756830/pdf/v053p00105.pdf

Dobosz, M., Lewandowski, M., Świerczewska, Z., Barańska-Rybak, W., & Cubała, W. J.(2023). Are YouTube videos a reliable source of information about body dysmorphic disorder? *Postep Psychiatr Neurol*. 32(2), 76-82. https://doi.org/10.5114/ppn.2023.128706 https://pmc.ncbi.nlm.nih.gov/articles/PMC10367510/

Drozd B, Couvillon E, Suarez A. (2018). Medical YouTube Videos and Methods of Evaluation: Literature Review. *JMIR Med Education*. 4(1). https://doi.org/10.2196/mededu.8527

Gimenez-Perez, G., Robert-Vila, N., Tomé-Guerreiro, M., Castells, I., & Mauricio, D. (2020). Are YouTube videos useful for

patient self-education in type 2 diabetes? *Health Informatics Journal.* 26(1), 45-55. https://doi.org/10.1177/1460458218813632

Global Web Index (GWI) (2023). Social Media behind the Scenes. https://www.gwi.com/webinars/social-media-behind-the-screens

Gulve, N., Tripathi, P., Dahivelkar, S., Gulve, M. Gulve, R., Kolhe, S. (2022). Evaluation of YouTube videos as a source of information about oral self-examination to detect oral cancer and precancerous lesions. *Journal of International Society of Preventive and Community Dentistry* 12(2) 226-234, https://doi.org/10.4103/jispcd.JISPCD_277_21

Hoffman, S. J., Conrad, K., & Benfield, J. (2019). YouTube as a source of health misinformation: An analysis of the quality of nutrition information available to youth. International *Journal of Environmental Research and Public Health*, 16(18), 3335. https://doi.org/10.3390/ijerph16183335

Horani, K., Coskey, A., & Hagedorn, J. C. (2023). Evaluating the quality of ankle fracture education on short-form video platform YouTube shorts. *Foot & Ankle Orthopaedics*.8(4). https://doi.org/10.1177/2473011423S00304

Iacopetta, D., Catalano, A., Ceramella, J., Pellegrino, M., Marra, M., Scali, E., Sinicropi, M. S., & Aquaro, S. (2024). The Ongoing Impact of COVID-19 on Pediatric Obesity. *Pediatric Reports*, 16(1), 135-150. https://doi.org/10.3390/pediatric16010013

Jang, E., Ko, E., Sim, J., Jeong, M. & Park, S. (2024). Mukbang media: correlations with the dietary behavior of children and adolescents. *Korea. Nutrition Research and Practice.* 18(5), 674-686. https://doi.org/10.4162/nrp.2024.18.5.674

Karabacak, M., & Margetis, K. (May 21, 2023) Embracing large language models for medical applications: opportunities and challenges. Cureus 15(5): e39305. https://doi.org/10.7759/cureus.39305

Khalil, M. Mohamed, F. & Shoufan, A. (2025). Evaluating the quality of medical content on YouTube using large language models. *Scientific Reports*, 15, (9906). https://doi.org/10.1038/s41598-025-94208-6

Kuchhal, D. & Li., F. (2022). A view into YouTube view fraud. In *Proceedings of the ACM Web Conference 2022* (WWW '22). Association for Computing Machinery, New York, NY, USA, 555–563. https://doi.org/10.1145/3485447.3512216

Larson, S. (2025, January 16). Social Media Users 2025 (Global Data & Statistics) https://prioridata.com/data/social-media-usage/

Lawrence, E., Mollborn, S., Goode, J., & Pampel, F. (2020). Health lifestyles and the transition to adulthood. *Socius*, 6. https://doi.org/10.1177/2378023120942070

Lee, Y., & Yoo, S. (2020). A study on the types and harmfulness of video ads after the change of content policy for children on YouTube. *The Korean Journal of Animation*, 16(3), 114-136. https://doi.org/10.51467/ASKO.2020.09.16.3.114

Leiner, M., de la Rosa, J. M., & de Vargas, C. (2025). The virtual kidnapping of youth by social media advertising*. Journal of Pediatric and Neonatal Individualized Medicine (JPNIM),* 14(1), e140105. https://doi.org/10.7363/140105

Malloy, J. A., Kemper, J. A., Partridge, S. R., & Roy, R. (2024). Empowering young women: A qualitative co-design study of a social media health promotion programme. *Nutrients*, 16(6), 780. https://doi.org/10.3390/nu16060780

Murthy, V. H. (2021). Confronting Health Misinformation: The U.S. Surgeon General's Advisory on Building a Healthy Information Environment https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf

Nuti, S. V., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R. P., Chen, S. I., & Murugiah, K. (2014). The use of Google Trends in health care research: A systematic review. *PLoS One,* 9(10), e109583. https://doi.org/10.1371/journal.pone.0109583

Obesity Action Coalition (2025). Learn about obesity - About obesity (causes and classifications). https://www.obesityaction.org/education-support/learn-about-obesity/causes/

Office of the surgeon General (2025). Health Misinformation. United States Department of Health and Human Services. https://www.hhs.gov/surgeongeneral/reports-and-publications/health-misinformation/index.html

Powell, L., Powell, G., Abdul, C., & Mariani, R. (2025). Tech for health: Understanding and identifying quality health information online. *Proceedings of the Northeast Decision Science Institute (NEDSI) Conference*. 296-300. https://nedsi.decisionsciences.org/wp-content/uploads/sites/5/2025/05/NEDSI-2025-Proceedings.pdf

Priftis, N., & Panagiotakos, D. (2023). Screen time and its health consequences in children and adolescents. *Children* (Basel). 10(10), 1665. https://doi.org/10.3390/children 10101665. PMID: 37892328.

Rajeshwari, B.S., Nayak, J.S., & Namratha, M. (2024). Impact on Fake News in Social Media and Current Technology in Detection of Fake News. In Kumar Soni, H., Sharma, S., & Sinha, G.R. (Eds.). Text and Social Media Analytics for Fake News and Hate Speech Detection (1st ed. Chapter 10). Chapman and Hall/CRC. https://doi.org/10.1201/978100 3409519

Son, R-H., Park, S. Y., Ko, Y-J., Jung, D. W., Won, E-S., Cho. A.H., Shin, D. H., & Kim, C-B. (2023). Community-Based intervention for elderly patients with diabetes mellitus: assessing the effectiveness of a YouTube self-education program during the global COVID-19 pandemic. *Korean Diabetes Association*. 24 (4), 232 -247. https://doi.org/10.4093/ jkd.2023.24.4.232

Sütcüoğlu, S., Özay, Z., I., Özet, A., Yazıcı, O., & Özdemir, N. (2023). Evaluation of scientific reliability and quality of YouTube videos on cancer and nutrition. *Nutrition*, 108, 111933. https://doi.org/10.1016/j.nut.2022.111933.

U.S. Centers for Disease Control and Prevention. (2024, April 2). Childhood Obesity Facts. U.S. https://www.cdc.gov/obesity/childhood-obesity-facts/childhood-obesity-facts.html

Vogels, E. A, Gelles-Watnick, R. & Massarata, R. (2022). Teens, Social Media and Technology 2022. Pew Research Center. https://www.jstor.org/stable/pdf/resrep63507.pdf?acceptTC=true&coverpage=false&addFooter=falseWang, M.L., Gago, C. M. (2024). Shifts in child health behaviors and obesity after COVID-19. *JAMA Pediatrics.* 178(5):427–428. https://doi.org/10.1001/jamapediatrics.2024.0027

Wartella, E., Rideout, V., Montague, H., Beaudoin-Ryan, L., & Lauricella, A. (2016). Teens, health and technology: a national survey. *Media and Communication*, 4(3), 13-23. https://doi.org/10.17645/mac.v4i3.515

Wijaya, L. S., & Subroto, U. (2025). The influence of fear of missing out on self-concept among high school students. AKADEMIK: *Journal Mahasiswa Humanis,* 5(1), 353–366. https://doi.org/10.37481/jmh.v5i1.1209

Xian Y, Lampert CH, Schiele B, & Akata Z (2019). Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 41(9), 2251-2265. https://doi.org/10.1109/TPAMI.2018.2857768

Appendix A

Table 3: Python Code for Fetching YouTube Videos with Parameters via a Google API

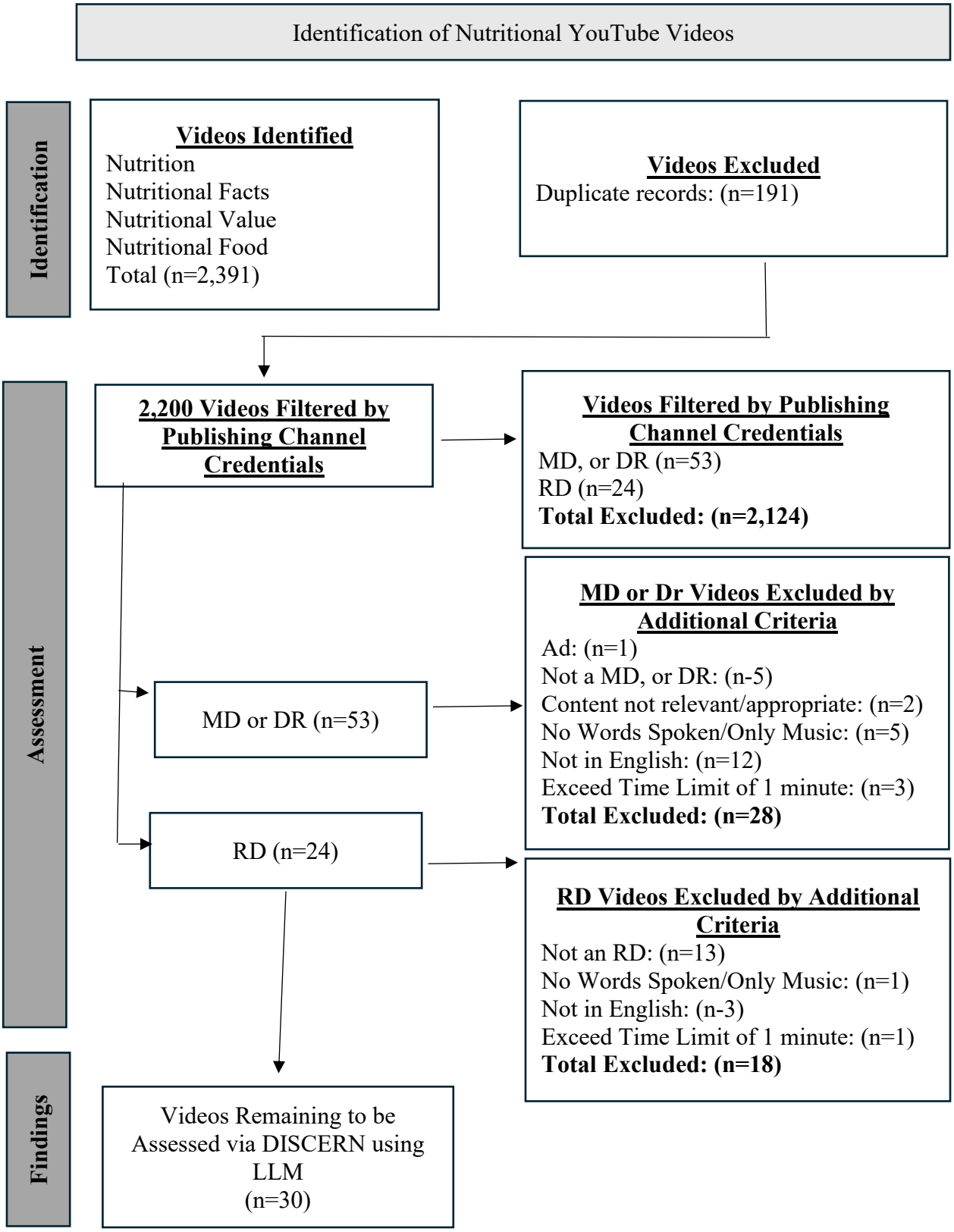
```
# First install isodate
pip install isodate

#After installing isodate, run the following code
import requests
from datetime import datetime
# To parse ISO 8601 durations
import isodate
# To export the results to a csv file, you will need to import csv
import csv
# The parameters of searching and pulling the youtube videos via Google API
def search_youtube_videos(api_key, search_queries, max_results_per_keyword=5000, output_csv="youtube_results.csv"):
    search_url = "https://www.googleapis.com/youtube/v3/search"
    video_details_url = "https://www.googleapis.com/youtube/v3/videos"
    start_date = "2023-01-01T00:00:00Z"
    end_date = "2025-01-01T00:00:00Z"
    all_videos = []
    for search_query in search_queries:
        # To exclude  ads use '-ad' in the search query
        refined_query = f"{search_query} -ad"
        params = {
            "part": "snippet",
            "q": refined_query,
            "type": "video",
        # To only fetch short videos  This is where Anaconda Clouds AI Assistance was used.
            "videoDuration": "short",
        # To fetch only those videos published after defined start date and end date
            "publishedAfter": start_date,
            "publishedBefore": end_date,
        # To fetch only those videos with a prioritize English-language
            "relevanceLanguage": "en",
            "maxResults": max_results_per_keyword,
            "key": api_key)
        next_page_token = None
        while True:
            if next_page_token:
                params["pageToken"] = next_page_token
            response = requests.get(search_url, params=params)
            if response.status_code != 200:
                print(f"Error fetching results for keyword '{search_query}':", response.json())
                break
            data = response.json()
            video_ids = [item["id"]["videoId"] for item in data.get("items", [])]
            if not video_ids:
                break
# To fetch video details (keywords, video title, channel title, URL, and duration in minutes)
            video_params = {
                "part": "contentDetails,snippet",
                "id": ",".join(video_ids),
                "key": api_key}
            video_response = requests.get(video_details_url, params=video_params)
            if video_response.status_code != 200:
```

```
            print(f"Error fetching video details for '{search_query}':", video_response.json())
            break
        video_data = video_response.json()
        for item in video_data.get("items", []):
            video_id = item["id"]
            title = item["snippet"]["title"]
            channel = item["snippet"]["channelTitle"]
            url = f"https://www.youtube.com/watch?v={video_id}"
            duration = isodate.parse_duration(item["contentDetails"]["duration"])
            duration_in_minutes = f"{int(duration.total_seconds() // 60)}:{int(duration.total_seconds() % 60):02}"
            all_videos.append({"keyword": search_query, "title": title, "channel": channel, "url": url, "duration":
duration_in_minutes})
        next_page_token = data.get("nextPageToken")
        if not next_page_token or len(all_videos) >= max_results_per_keyword:
            break
# To export or write results to CSV file
    with open(output_csv, mode="w", newline="", encoding="utf-8") as file:
        writer = csv.DictWriter(file, fieldnames=["keyword", "title", "channel", "url", "duration"])
        writer.writeheader()
        writer.writerows(all_videos)
    print(f"Results have been saved to {output_csv}")
# We removed our API Key and replaced it with  X's
# Replace the Xs with your API key
api_key = "XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX"
# To fetch videos via our 4 search terms
search_queries = ["Nutrition", "Nutritional Facts", "Nutritional Food", "Nutritional Value"]
# To print the results to an output file titled "my_results.csv"
search_youtube_videos(api_key, search_queries, max_results_per_keyword=5000, output_csv="my_results.csv")
```

Appendix B

PRISMA flow diagram for video selection process



Appendix C

Table 4: Details Regarding the Included Shorts

Short	Length	No. of Subscribers	No. of Views	No. of Likes	No. of Comments	Country	Cross Channeled on Instagram or TikTok	Credential
1	0.042	172,000	3,765	58	33	USA	YES	MD
2	0.041	493,000	72,327	5,200	260	USA	YES	MD
3	0.024	6,460	306	7	0	USA	YES	MD
4	0.032	31,000	14,980	497	3	USA	YES	DR
5	0.041	38,600	1,613	47	1	India	YES	DR
6	0.04	226	120	5		USA	NO	DR
7	0.031	24,000	149,606	3,500	66	Germany	NO	DR
8	0.035	24,000	4,605	234	5	Germany	NO	DR
9	0.04	523,000	387,765	17000	241	USA	YES	MD
10	0.04	523,000	5,921	360	13	USA	YES	MD
11	0.02	523,000	17,981	858	6	USA	YES	MD
12	0.023	523,000	4,501	206	1	USA	YES	MD
13	0.013	3,530,000	419,364	15000	453	USA	NO	MD (Organization)
14	0.017	3,530,000	384,694	16000	1006	USA	NO	MD (Organization)
15	0.016	3,530,000	63,545	4000	306	USA	NO	MD (Organization)
16	0.022	3,530,000	284,899	20000	563	USA	NO	MD (Organization)
17	0.026	3,530,000	590,522	35000	831	USA	NO	MD (Organization)
18	0.014	3,530,000	251,959	12000	362	USA	NO	MD (Organization)
19	0.039	3,530,000	1,072,971	66000	1543	USA	NO	MD (Organization)
20	0.026	3,530,000	42,851	4500	230	USA	NO	MD (Organization)
21	0.037	1,790,000	21,388,509	1300000	4591	USA	YES	RD
22	0.039	1,790,000	6,836,922	458000	2654	USA	YES	RD
23	0.035	1,790,000	10,917,808	614000	2651	USA	YES	RD
24	0.017	1,790,000	1,395,474	105000	348	USA	YES	RD
25	0.041	1,790,000	6,680,914	413000	2570	USA	YES	RD
26	0.028	1,790,000	1,768,767	133000	719	USA	YES	RD
27	0.032	14,100	584,720	17000	125	USA	NO	MD (Organization)
28	0.042	1,430,000	37,943	3300	81	USA	YES	MD
29	0.026	81,600	46,781	1900	72	USA	YES	MD
30	0.019	351,000	6,687	579	13	USA	NO	MD

Appendix D

Table 5: DISCERN Scores by Short and Question
(produced by Gemini 2.5 Flash Preview 05-20)

	Reliability									Quality of Information on Treatment Choices								Overall Quality
Video No.	Q 1	Q 2	Q 3	Q 4	Q 5	Q 6	Q 7	Q 8	Overall Reliability Score	Q 9	Q 10	Q 11	Q 12	Q 13	Q 14	Q 15	Overall Quality of Info on Treatment Choices	Q 16
1	3	3	5	1	1	1	1	1	2.00	3	3	3	3	3	1	1	2.43	1.00
2	5	5	5	3	3	3	1	3	3.50	3	3	5	3	1	3	1	2.71	3.00
3	3	3	5	1	1	3	1	1	2.25	1	1	1	1	1	3	1	1.29	1.00
4	5	5	5	1	1	3	1	1	2.75	3	5	1	1	1	1	1	1.86	3.00
5	5	5	3	3	1	1	1	1	2.50	3	5	1	1	3	1	1	2.14	1.00
6	5	5	5	3	3	3	1	1	3.25	3	3	5	3	1	3	1	2.71	3.00
7	5	5	5	1	1	3	1	1	2.75	5	3	1	1	3	1	1	2.14	3.00
8	5	3	5	1	1	1	1	1	2.25	1	3	3	3	3	1	1	2.14	1.00
9	5	5	5	1	1	3	1	1	2.75	5	3	1	1	1	1	1	1.86	3.00
10	5	5	3	1	1	1	1	1	2.25	3	5	1	1	1	1	1	1.86	1.00
11	5	5	5	1	1	3	1	1	2.75	3	3	3	3	3	1	1	2.43	3.00
12	3	1	1	1	1	1	1	1	1.25	1	1	1	1	1	1	1	1.00	1.00
13	5	5	3	1	1	1	1	1	2.25	1	5	1	3	1	1	1	1.86	1.00
14	5	5	5	1	1	3	1	1	2.75	5	1	3	1	1	3	1	2.14	3.00
15	5	5	5	1	1	3	1	1	2.75	1	3	1	1	1	3	1	1.57	3.00
16	5	5	5	1	1	1	1	1	2.50	3	1	1	1	1	1	1	1.29	1.00
17	5	5	5	1	1	1	1	1	2.50	3	5	1	1	3	1	1	2.14	1.00
18	5	5	3	1	1	1	1	1	2.25	5	3	1	1	1	1	1	1.86	1.00
19	5	5	5	1	1	1	1	1	2.50	3	3	1	3	1	1	1	1.86	1.00
20	5	5	3	1	1	1	1	1	2.25	1	1	1	1	1	1	1	1.00	1.00
21	5	5	5	3	1	5	1	3	3.50	5	5	3	3	5	3	3	3.86	5.00
22	5	5	5	3	1	5	1	3	3.50	5	5	3	1	5	3	3	3.57	5.00
23	5	5	5	3	1	5	1	3	3.50	5	5	3	1	5	3	3	3.57	5.00
24	5	5	5	3	1	5	1	3	3.50	5	5	3	1	3	3	3	3.29	5.00
25	5	5	5	3	1	5	1	3	3.50	5	5	5	5	3	3	3	4.14	5.00
26	5	5	5	3	1	5	1	3	3.50	5	5	3	1	5	3	3	3.57	5.00
27	5	5	5	1	1	3	1	1	2.75	3	5	1	1	3	1	1	2.14	3.00
28	5	5	5	1	1	3	1	1	2.75	3	5	3	3	5	3	1	3.29	3.00
29	5	5	5	1	1	1	1	1	2.50	5	3	1	1	1	1	1	1.86	1.00
30	5	5	3	1	1	1	1	1	2.25	3	3	1	1	1	1	1	1.57	1.00

Data-Driven Peer Group Selection for Salary Comparison in Higher Education: An Applied Analytics Approach to Building Trust

Eric Allen Breimer
ebreimer@siena.edu
Siena University
Loudonville, NY 12211

Sangahn Kim
skim@siena.edu
Siena University
Loudonville, NY 12211

Seung Jin Wang
swang@siena.edu
Siena University
Loudonville, NY 12211

Abstract

This study introduces a data-driven method to select peer institutions in higher education for faculty salary comparison. Given a target institution, the goal is to form a peer group of similar colleges using stakeholder-identified variables like enrollment, finances, and student outcomes, but excluding salary data. An effective peer group places the target near the median salary level. Previous work raised equity concerns because the methodology generated separate peer groups, including one for base salaries and several for high-demand accredited disciplines. Concerns about objectivity and fairness emerged due to the use of subjective filters and post-hoc adjustments, such as including aspirational institutions. We seek a more consistent, data-driven approach that uses principal component analysis and nearest neighbor search to create a single unified peer group that can be used for all salary comparisons. By employing a more transparent, analytics-based method, we aim to enhance trust in the process to promote acceptance of the peer group among faculty and administrative stakeholders.

Keywords: Data-driven methodology, peer institution selection, salary benchmarking, higher education compensation, principal component analysis, nearest neighbor analysis.

Recommended Citation: Breimer, E.A., Kim, S., Wang, S., (2026). Data-Driven Peer Group Selection for Salary Comparison in Higher Education: An Applied Analytics Approach to Building Trust. *Journal of Information Systems Applied Research and Analytics*, v19(n4) pp 70-82. DOI# https://doi.org/10.62273/HGCB4142

Data-Driven Peer Group Selection for Salary Comparison in Higher Education: An Applied Analytics Approach to Building Trust

Eric Allen Breimer, Sangahn Kim Seung Jin Wang

1. INTRODUCTION

In higher education, equitable compensation for fulltime faculty is a cornerstone of institutional stability and morale. Compensation models often rely on benchmarking against peer institutions to ensure comparability, accounting for variations across disciplines where market forces may drive salary differentials. Traditionally, institutions have employed multiple peer groups tailored to specific purposes: a base group for general salaries and specialized groups for high-demand accredited fields that often have significantly higher salaries.

A fragmented approach to peer group selection can create perceptions of inequity among faculty and administrative stakeholders. For example, when peer groups for different disciplines vary significantly in composition, disparities in financial metrics may raise concerns about fairness. Additionally, common practices such as ad hoc filtering to exclude problematic peers and the subjective inclusion of aspirational institutions often lack consistent, rigorous criteria, undermining trust in the process. When faculty and administrators lack confidence in the peer group selection, it becomes challenging to accept compensation decisions based on comparisons with those groups.

This paper presents a data-driven methodology termed the Unified Peer Group (UPG), designed to streamline peer selection into a single, consistent framework. The UPG combines the top overall nearest neighbors with the most similar peers that share discipline-specific accreditations with the target institution. This approach enables the entire UPG to guide base salary decisions, while subsets can inform adjustments for accredited high-demand disciplines.

By leveraging Principal Component Analysis (PCA) and nearest neighbor analysis, the UPG identifies institutions most similar to the target institution across a multidimensional space. Salary data are excluded from the peer selection process to ensure impartiality. The objective is to form a peer group where the target institution aligns near the median for multiple variables/features. This approach assumes that the target's salaries will similarly approximate the peer group's median. If this assumption fails, it indicates that the target's salaries deviate from those of comparable peers, potentially justifying compensation adjustments.

Our new methodology improves upon prior work from 2021 and 2024 where subjective filters on Carnegie classification (American Council on Education, 2025), public/private status, and geography were deemed essential.

Empirical evidence shows that the refined, data-driven selection process efficiently excludes unsuitable peers, minimizing the need for ad hoc filters. Although administrative stakeholders recommended retaining one filter, this methodology has significantly strengthened trust in the peer group and decision-making process at the authors' institution.

2. BACKGROUND

Peer institution selection is a critical process in institutional research, evolving from subjective, bias-prone methods to sophisticated, data-driven approaches. Early peer selection relied on subjective criteria like geographic proximity or mission alignment, which often introduced inconsistencies (D'Allegro, 2017; D'Allegro & Zhou, 2013). These studies highlight the limitations of such approaches, advocating for objective methodologies using Integrated Postsecondary Education Data System (IPEDS) data. For instance, McLaughlin et al. (2011) proposed nearest neighbor algorithms to form peer groups based on key institutional metrics such as enrollment, finances, and student outcomes, offering a reproducible framework that minimizes bias.

To enhance the precision of peer selection, advanced analytical techniques like Principal Component Analysis (PCA) have gained prominence. PCA reduces correlated variables into uncorrelated principal components, capturing essential data variance while simplifying complex datasets (Jolliffe & Cadima, 2016). When integrated with nearest neighbor algorithms, PCA improves classification

accuracy, as demonstrated in educational and non-educational contexts (Lubis et al., 2020; McLaughlin et al., 2011). This synergy of PCA and nearest neighbor methods provides a robust foundation for equitable peer comparisons, particularly in contexts like salary benchmarking, where fairness and transparency are paramount.

In higher education, peer benchmarking informs critical policy decisions, such as funding and compensation strategies (Kelchen et al., 2024). However, existing multi-group models often fail to account for contextual factors like accreditation, which can significantly influence institutional profiles (AACSB, 2025; CCNE, 2025). Recent analytics applications in institutional research, such as big data in healthcare and campus crime analysis, underscore the importance of transparent, data-driven frameworks to build trust and ensure equitable comparisons (Mohammed & Lind, 2024; Kline et al., 2020). Despite these advances, gaps remain in integrating categorical factors like accreditation into unified peer selection models.

This study addresses these gaps by proposing a novel framework that combines PCA, nearest neighbor algorithms, and accreditation as a categorical factor. By synthesizing data-driven methodologies (McLaughlin et al., 2011; Lubis et al., 2020) with contextual considerations (AACSB, 2025; CCNE, 2025), this approach aims to enhance the accuracy and equity of salary benchmarking, contributing to more informed resource allocation in higher education.

3. METHODS

Data Sources and Variables

In our work, data were downloaded from IPEDS (NCES, 2023) focusing on 2,605 institutions with sufficient reported data (at least 11 of 14 key columns). Missing values were imputed using nearest neighbors following the approach of Troyanskaya et al. (2001). The authors' home institution has two key accreditations that impact salaries, Business (AACSB, 2025) and Nursing (CCNE, 2025). An important stakeholder goal was to include a balanced mix of peers with AACSB accreditation, CCNE accreditation, both accreditations, and neither. The accreditation status of institutions in not included in IPEDS and was scraped directly from the AACSB and CCNE websites.

Stakeholders expressed concerns about deviating significantly from past approaches, emphasizing the need for year-to-year consistency. Our goal was to introduce a new peer selection methodology that would gain broad acceptance without significantly altering the core variable set, which could raise additional concerns. Once a methodology is adopted, future work can explore refinements to variable selection.

The key variables shown below were derived from 14 IPEDS columns (NCES, 2023) and direct accreditation data. These variables seek to capture institutional size (student and faculty counts), financial health, and student success metrics, aligning with past practices. We introduced one new variable to capture key accreditations. We excluded Carnegie classification, geographic location, and institutional type (public vs. private) which were the subjects of subjective ad hoc filtering that previously complicated peer selection.

FTEGD: Full-time equivalent graduate students.

FTEUG: Full-time equivalent undergraduates

Revenue: Total operating revenue

Endowment: Value of the endowment

Net Assets: Total assets including endowment

Ret Rate: Retention rate from year 1 to 2

Grad Rate: Four-year graduation rate.

Adm Rate: Students enrolled divided by students admitted

Faculty FTE: Full-time equivalent faculty, weighted by full/part-time

Net Price: Average price after discounts, weighted by graduate/undergraduate

ACCRED: Accreditation

An accreditation (ACCRED) value of 1 indicates a school that shares all the key accreditations of the target institution, the value 0 indicates no shared accreditations, and the intermediate values indicate the percentage of shared accreditations. For example, given a target institution with five key accreditations, a peer that shares 4 out of the 5 accreditations would have an ACCRED value of 0.8. Representing all accreditations as one column helps to avoid the over-weighting that might occur when considering many key accreditations stored as separate variables. While this paper focuses on the authors' home institution with two key accreditations, our methodology can scale for institutions with many key accreditations.

Analytical Process

PCA was applied after normalizing the variables with standard scaling (Jolliffe, 2002). The top 9 components, explaining 99% of the variance, were selected to ensure the ACCRED variable, which had low weight in components 1–8, influences peer selection. Figure 1 shows the cumulative explained variance of the principal components. Figure 2 shows the loadings or weighting of each of the direct variables on the first five principal components (see Appendix A for the full table).

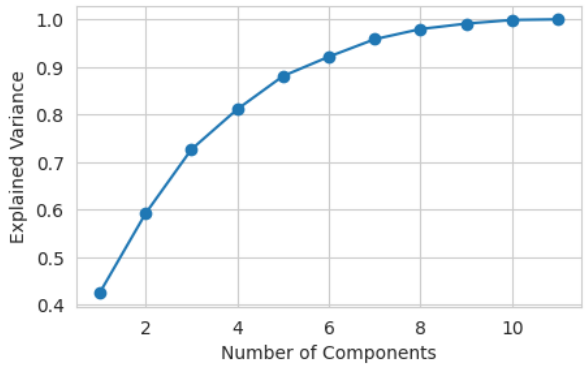


Figure 1: Cumulative explained variance by PCA components

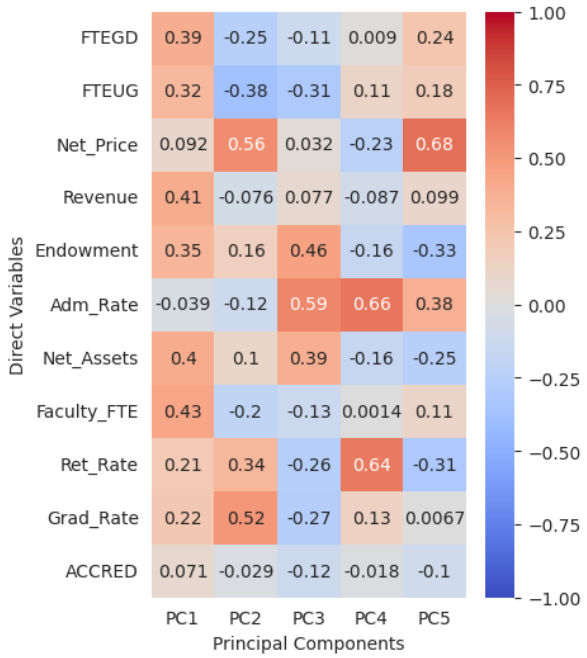


Figure 2: Loadings (weighting) of direct variables on first 5 principal components

Nearest neighbors were computed using Gower distance (Gower, 1927) rather than Euclidean distance, due the presence of the ACCRED variable. Gower distance is considered a good choice given categorical variables. While the ACCRED variable could be considered the percentage of matched accreditations, it exhibits characteristics of an ordinal categorical value when considering on a few key accreditations. Additionally, IPEDS includes some variables with small value ranges that are rounded, which also have categorical characteristics. Thus, we felt Gower distance was the best choice given the composition of variables.

The Unified Peer Group (UPG) concept combines the top overall nearest neighbors with the top accredited neighbors for each key accreditation. This method ensures sufficient accredited peers to provide robust salary data for high-demand disciplines while maintaining a manageable UPG size for stakeholder review. At the authors' institution, stakeholders deemed a UPG exceeding 50 institutions too large and fewer than 15 insufficient for reliable salary data. Historically, peer groups ranged from 15 to 30 institutions, and significant deviations from this range raised concerns about reduced stakeholder acceptance.

To account for stakeholder concerns, the authors' institution defined the UPG as the union of:

1. Top 30 overall nearest neighbors.
2. Top 15 AACSB-accredited neighbors.
3. Top 15 CCNE-accredited neighbors.
4. Additional non-accredited neighbors to ensure at least 33% of the UPG lacks either accreditation.

Although the specific number of schools in the UPG is not determined through data-driven methods, the size selection aims to align with historical peer groups to enhance stakeholder acceptance. In general, the UPG definition should vary based on stakeholder concerns and constraints at the target institution.

In the general case, it important to note that the intersection of top overall nearest neighbors and accredited neighbors may vary. In one extreme, the top overall peers may include no accredited institutions, but the UPG definition ensures a minimum number of peers for each key accreditation. Conversely, if the top overall peers hold all key accreditations, stakeholders may raise concerns about their over-representation. However, the UPG definition can be adjusted to ensure a minimum number of non-accredited peers.

At the authors' institution, stakeholders

recommended that one-third of the peer group consist of schools without either accreditation to reflect historical peer group composition. This proportion can be set to zero for institutions whose peer groups historically consisted entirely of accredited institutions.

Although the selection of UPG size and composition involves inherent subjectivity, the UPG framework establishes overarching goals rather than ad hoc procedures for specific institution selection. For instance, past practices—such as excluding schools based on Carnegie classification—served as tactical steps to refine the peer list, not as strategic aims for achieving a particular Carnegie composition. Our work emphasizes developing an unbiased, data-driven process for selecting peers within stakeholder-defined parameters, leaving size determination to consensus. While our goal was to implement a fully data-driven approach, stakeholder feedback at the author's institution necessitated one subjective post-hoc filter: the exclusion of doctoral institutions.

Finally, the entire process is implemented in Python and documented in a Google Colab notebook (Google, n.d.). The notebook includes data acquisition (downloading and scraping), cleaning, principal component analysis, nearest neighbor calculations, and supporting visualizations. The notebook serves as an audit trail, enabling stakeholders to thoroughly examine the process.

4. RESULTS & ANALYSIS

Applying PCA and nearest neighbor yielded a ranking of the 2,605 institutions that we considered. The distribution of the distances to the authors' home institution (the target) are shown in Appendix B. Appendix C shows the correlation of key variables and the targets position in the distribution of key variables. We repeated the analysis for 20 random targets. This section includes a summary focused on geographic proximity, public vs private, Carnegie class and accreditation.

Geographic Proximity

In the past, geographic filters were an important element in peer group selection. Stakeholders felt that schools in distant regions would be poor matches and should be filtered out. However, such filtering suffers from subjectivity, especially bias in defining the boundaries of the target region. For instance, restricting a peer group to Northeastern states might exclude Ohio, even though Ohio may exhibit substantial similarity to the target region. Thus, the exclusion of Ohio-based schools might represent an ineffective filter.

Our results indicated that geographic filtering may not be necessary. Figure 3 shows the geographic clustering of peer groups from three randomly selected target institutions. The other randomly selected targets exhibited similar clustering where peers tend to be geographically closer to the target institution.

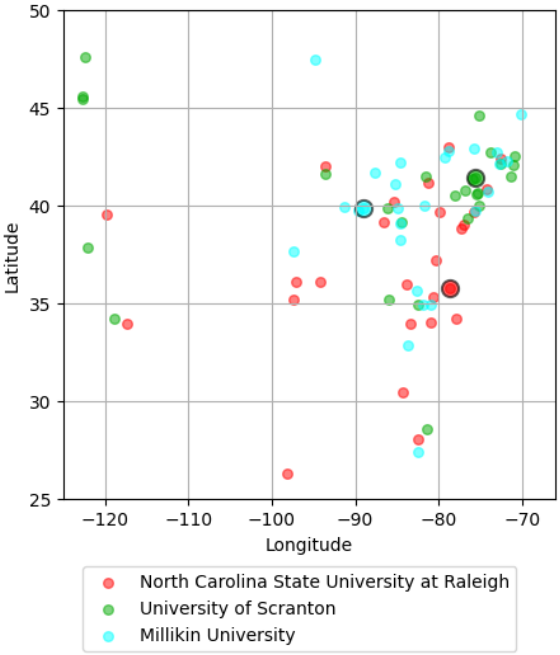


Figure 3: Geographic distribution of three example peer groups

Figure 4 shows a breakdown of Scranton University in Pennsylvania (PA) and Millikin University in Illinois (IL). The 10 geographically closest states to the targets are shown in green highlighting the tendency for peers to be in the nearest states.

There are many reasons institutions that are similar in key variables might be geographically close. Economically prosperous regions can support types of institutions that other regions cannot support. Since many students attend colleges near home (Turley, 2009; Acton, 2024), local schools compete to attract the same cohorts. Thus, nearby schools may converge in student quality similarity.

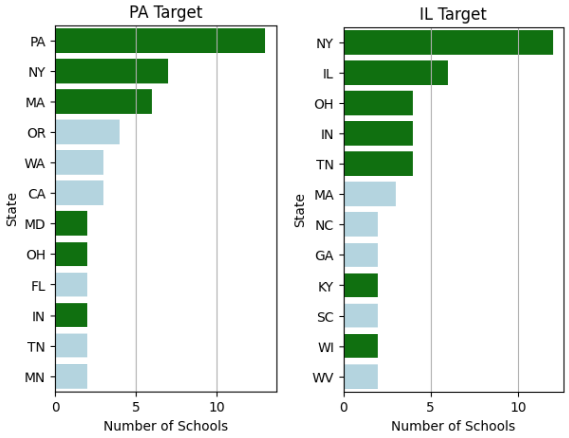


Figure 4: State distribution of peer groups for target schools in PA and IL

At the authors' home institution in the Northeast, stakeholders were initially skeptical when schools far outside the region were selected as nearest neighbors. However, after researching these geographic outliers, stakeholder agreed that they were good selections. Stakeholder were also comfortable including a few geographic outliers as long as the majority of peers were within the region. Using Nearest Neighbor without any geographic filtering is an opportunity to discover excellent matches outside of the region. And, the tendency to select peers in the region gives stakeholder confidence in the overall process.

Public vs Private

The authors' home institution is a private 4-year college and stakeholders felt that filtering out public institutions was essential. However, only 3 public institutions ranked among the top 200 nearest neighbors (ranked #185, 187 and 198, respectively). When key financial variables are included, public and private institutions demonstrate significant difference. This gave stakeholder further confidence in the nearest neighbor approach in selecting appropriate peers.

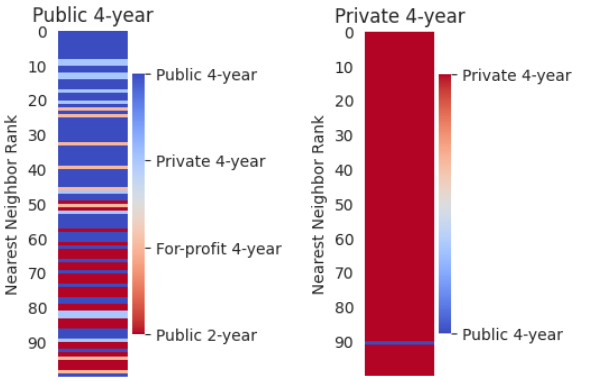


Figure 5: Visualization of nearest neighbors of a target public 4-year and private 4-year.

Figure 5 visualizes the 100 nearest neighbors of a public 4-year institution and a private 4-year institution. For the public 4-year target, 32 out of the top 50 peers were also public 4-year institutions. Other public 4-year targets exhibited similar distributions. Private four-year institutions are the most common among the 2,605 schools considered. As a result, more private schools are available for selection, and public institutions rarely rank among the top peers of a private target institution.

Carnegie Class

In 2021, the authors' home institution in the Northeast region was reclassified from *Baccalaureate Colleges: Arts & Sciences Focus* to *Master's Colleges & Universities: Smaller Programs*. This reclassification stemmed from earning AACSB accreditation in 2007, CCNE accreditation in 2017, launching a Master of Science in Accounting in 2009, and introducing an MBA program in 2018. This context is very important because some institutions may on the edge between two Carnegie classes or may overlap with two or more classes.

In the past, peer groups were selected by considering institutions that matched the target's current and most recent previous Carnegie classification. This decision was made without investigating the similarity of the target to schools in other Carnegie classes. Thus, this filter yielded very few peers with either AACSB or CCNE accreditation. Afterwards, stakeholder would advocate for the inclusion of accredited aspirants, which was an inherently subjective process.

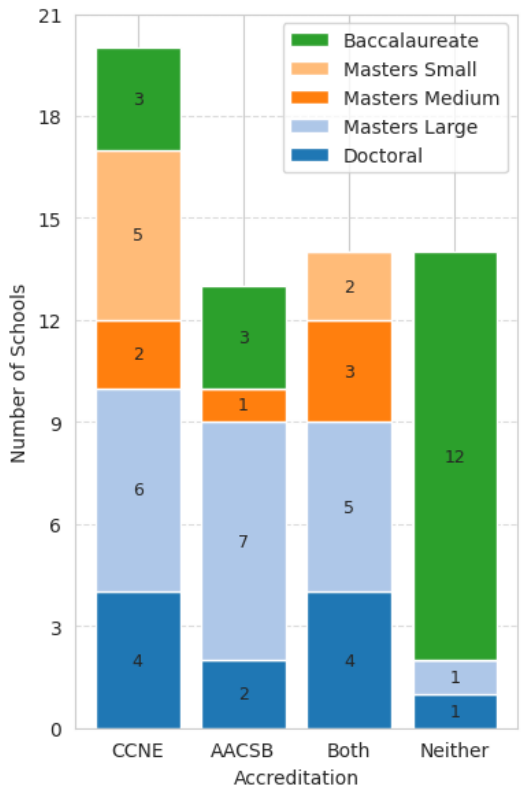


Figure 6: Carnegie breakdown of the top 60 nearest neighbors to the target.

Figure 6 shows the top 60 nearest neighbors to the authors' home institution, which includes (a) 20 schools with only CCNE accreditation, (b) 13 schools with only AACSB accreditation, (c) 14 schools with both accreditations, and (d) 14 with neither.

In examining these four groups, it became clear that peers in Carnegie class *Large and Medium Master's Colleges & Universities* should be selected in order to form a group with a sufficient number of accredited peers (at least 15 of each).

Nearest neighbor revealed that many of the closest matched accredited peers were outside of the target's Carnegie class. Examining these peers revealed significant similarities, particularly in overall enrollment and financial variables. Although the Carnegie size classification reflects graduate program size—and the target institution has relatively smaller graduate enrollment—there was strong alignment in tuition-based revenue and weighted cost. Stakeholders ultimately concluded that including medium and large master's institutions was appropriate, especially given the institution's strategic goal of increasing graduate enrollment.

Carnegie *Doctoral/Professional Universities* (R3), ranked among the top 60 at positions 6, 9, 20, 21, 24, 26, 38, 42, 52, 54, and 60. Further investigation revealed that these nearest neighbors had only a few small doctoral programs. However, since the target institution has no plans to start any doctoral programs, stakeholders felt that peers with doctoral programs should be excluded from the UPG. This was the only post-hoc filter applied, and it was widely accepted by stakeholders.

The Unified Peer Group

After excluding doctoral institutions and applying the union criteria, the UPG consisted of 36 institutions with the following characteristics:

- 7 with both accreditations.
- 12 with neither (33% threshold met).
- 16 CCNE total (9 CCNE-only + 7 both).
- 15 AACSB total (8 AACSB-only + 7 both).

Note that the 15th selected AACSB peer is also CCNE-accredited.

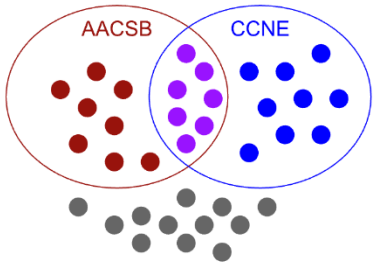


Figure 7: Unified Peer Group (UPG) overlap

Figure 7 illustrates the overlap among the 24 schools with one or both accreditations. This overlap—including the 7 schools with both accreditations—results in a relatively small peer group that captures sufficient CCNE and AACSB peers for discipline-specific salary comparisons.

To understand the influence of accreditation on selection, we re-ran PCA and nearest neighbor without the ACCRED variable and it yielded a nearly identical UPG of 36 institutions. In this new UPG, the lowest ranked peer with both accreditations was dropped in favor of a better match with only AACSB accreditation. With ACCRED excluded, the 15th-ranked CCNE and AACSB peers were ranked 39th and 43rd respective out of 2,605 total institutions (874 with CCNE and 554 with AACSB). Among the top 30 nearest neighbors were 9 AACSB peers and 13 CCNE peers.

5. DISCUSSION & CONCLUSION

The overall goal of this work was to establish a peer group methodology that could be widely accepted by stakeholders. To achieve this goal, we considered two related objectives. First, we aimed to establish a single unified peer group that could be used to determine base salaries for all faculty, as well as discipline-specific salaries for accredited programs. Second, we sought to develop a data-driven process that eliminated as many subjective filters, ad hoc processes, and post-hoc decisions as possible.

In the past, independent peer groups were developed for discipline-specific salary comparisons, dividing the institution. For instance, the base group and Nursing group were selected with Carnegie filters that excluded most institutions with graduate programs, whereas the Business group, out of necessity, primarily included institutions with medium and large master's programs.

The UPG mitigates institutional division, equity concerns and stakeholder mistrust in two key ways. First, the selection process is consistent for all accredited discipline-specific groups. Second, base salaries are influenced by the inclusion of accredited peers in the UPG.

When a target institution holds multiple key accreditations, a key challenge is preventing the UPG from becoming excessively large. Stakeholders often seek to investigate all peers more deeply to understand each selection, which is not practical with a very large UPG. However, the UPG must include enough peers per accredited discipline to enable robust salary comparisons. The UPG framework allows adjustments to achieve a manageable size, such as modifying the number of peers per accreditation or the proportion of non-accredited peers. Thus, the UPG provides a flexible, scalable framework that can be adjusted to the institution's accreditations and stakeholder needs.

Our approach demonstrates empirically that certain contentious subjective filters may be unnecessary. For instance, among the top 200 nearest neighbors to a private, small Master's institution, only three were public institutions. Moreover, the closest Carnegie R1 university ranked 136th. These results indicate that nearest neighbor analysis naturally excludes unsuitable peers from the top matches. Furthermore, the data-driven process revealed that geographic and Carnegie classification filters excluded some of the best-matched peers, as observed by stakeholders. Our data-driven method enabled stakeholders to recognize that excellent matches often included institutions beyond the target's geographic region or Carnegie classification.

At the authors' institution, the new methodology increased trust, consistent with literature on analytics adoption (Mohammed & Lind, 2024). Specifically, a unified, data-driven peer selection process—using PCA and nearest neighbors—produced a peer group that was quickly accepted with only one post-hoc adjustment: filtering out doctoral universities. Previously, post-hoc changes, such as including aspirational institutions, required subjective and time-consuming negotiation among faculty and administrative stakeholders.

6. FUTURE WORK

Our analysis revealed that accredited peers are often selected as nearest neighbors, even when accreditation is excluded as an input variable. However, this finding is based on a single target case, and further analysis is needed to assess accreditation's direct impact on peer selection. Accreditation likely correlates with other variables, leading to the selection of accredited peers due to their similarity in other features.

With confidence in the new methodology, a broader range of variables can be considered for updating the peer group in future years. Sources like IPEDS provide hundreds of variables, making comprehensive feature analysis a logical next step for refining the process. Past methodologies raised concerns about overfitting and weighting bias when using numerous correlated variables. By employing PCA, our approach enables the inclusion of a broader range of variables to assess similarity with the target institution.

At the time of this writing, administrative stakeholders are examining the salaries of institutions in the UPG. If the target institution is near the median of the UPG then our methodology will be further validated.

9. REFERENCES

AACSB (2025). Association to Advance Collegiate Schools of Business. (n.d.). AACSB International. https://www.aacsb.edu/

Acton, R., Cortes, K. E., Miller, L., & Morales, C. (2024). Distance to degrees: How college proximity shapes students' enrollment choices and attainment across race-ethnicity and socioeconomic status (EdWorkingPaper No. 24-1055). Annenberg Institute at Brown University. https://doi.org/10.26300/vjyg-ta27

American Council on Education (2025). Carnegie Classifications. https://carnegieclassifications.acenet.edu/

CCNE (2025). Commission on Collegiate Nursing Education. (n.d.). CCNE accreditation. American Association of Colleges of Nursing. https://www.aacnnursing.org/ccne-accreditation

D'Allegro, M. L. (2017). A case study to examine three peer grouping methodologies. The AIR Professional File, (142). https://doi.org/10.34315/apf1422017

D'Allegro, M. L., & Zhou, K. (2013). A case study to examine peer grouping and aspirant selection. The AIR Professional File, (132). https://doi.org/10.34315/apf1322013

Google. (n.d.). Google Colab (Version 3.0) [Software]. https://colab.research.google.com

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27(4), 857–871.

Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065). https://doi.org/10.1098/rsta.2015.0202

Kelchen, R., Ortagus, J., Rosinger, K., Baker, D., & Lingo, M. (2024). The relationships between state higher education funding strategies and college access and success. Educational Researcher, 53(2), 100–110. https://doi.org/10.3102/0013189X231208964

Kline, D., Vetter, R., Clark, U., (2020). Understanding Campus Crime with a Multi-University Analytics System. *Journal of Information Systems Applied Research*13(3) pp 21-28.

Lubis, A. H., Sihombing, P., & Nababan, E. B. (2020). Analysis of accuracy improvement in K-nearest neighbor using principal component analysis (PCA). Journal of Physics: Conference Series, 1566, 012062. https://doi.org/10.1088/1742-6596/1566/1/012062

McLaughlin, G. W., Howard, R. D., & McLaughlin, J. (2011, May 21–25). Forming and using peer groups based on nearest neighbors with IPEDS data [Paper presentation]. 51st Annual Forum of the Association for Institutional Research, Toronto, Ontario, Canada. https://eric.ed.gov/?id=ED504414

Mohammed, A., Lind, M., (2024). Examining Factors Influencing the Acceptance of Big Data Analytics in Healthcare. Journal of Information Systems Applied Research 17(2) pp 31-44. https://doi.org/10.62273/QNDU3179

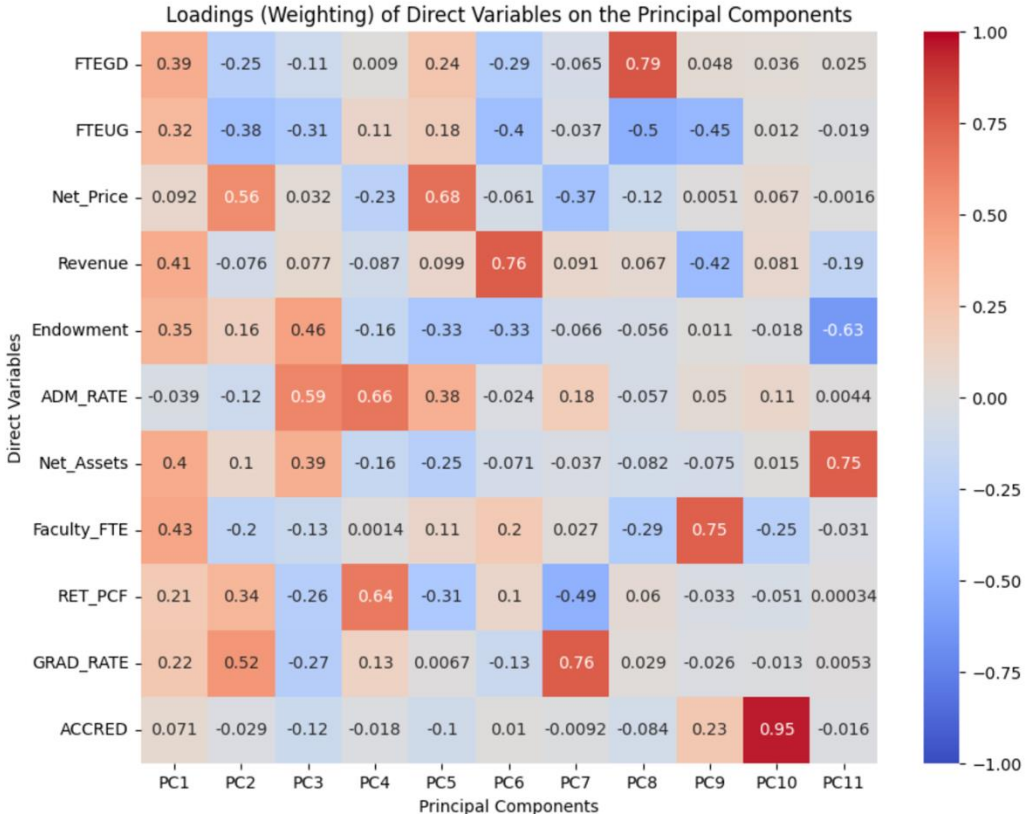
NCES (2023). National Center for Education Statistics. Integrated Postsecondary Education Data System (IPEDS), 2022–23 final data. U.S. Department of Education. https://nces.ed.gov/ipeds/use-the-data/download-access-database

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

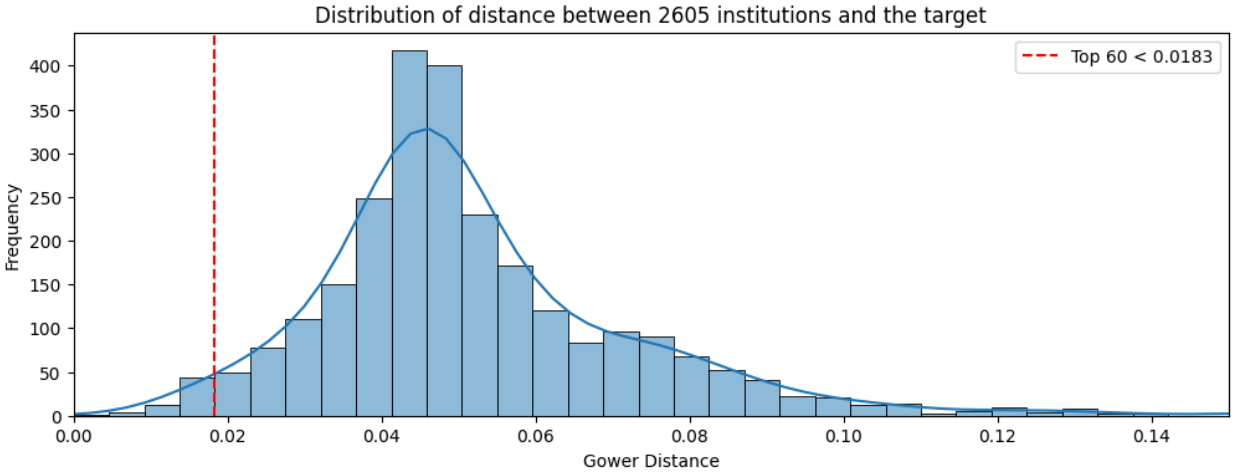
Turley, R. N. L. (2009). College proximity and the urban isolation of urban youth. Social Science Research, 38(3), 628–646. https://doi.org/10.1016/j.ssresearch.2009.02.002

Appendices and Annexures

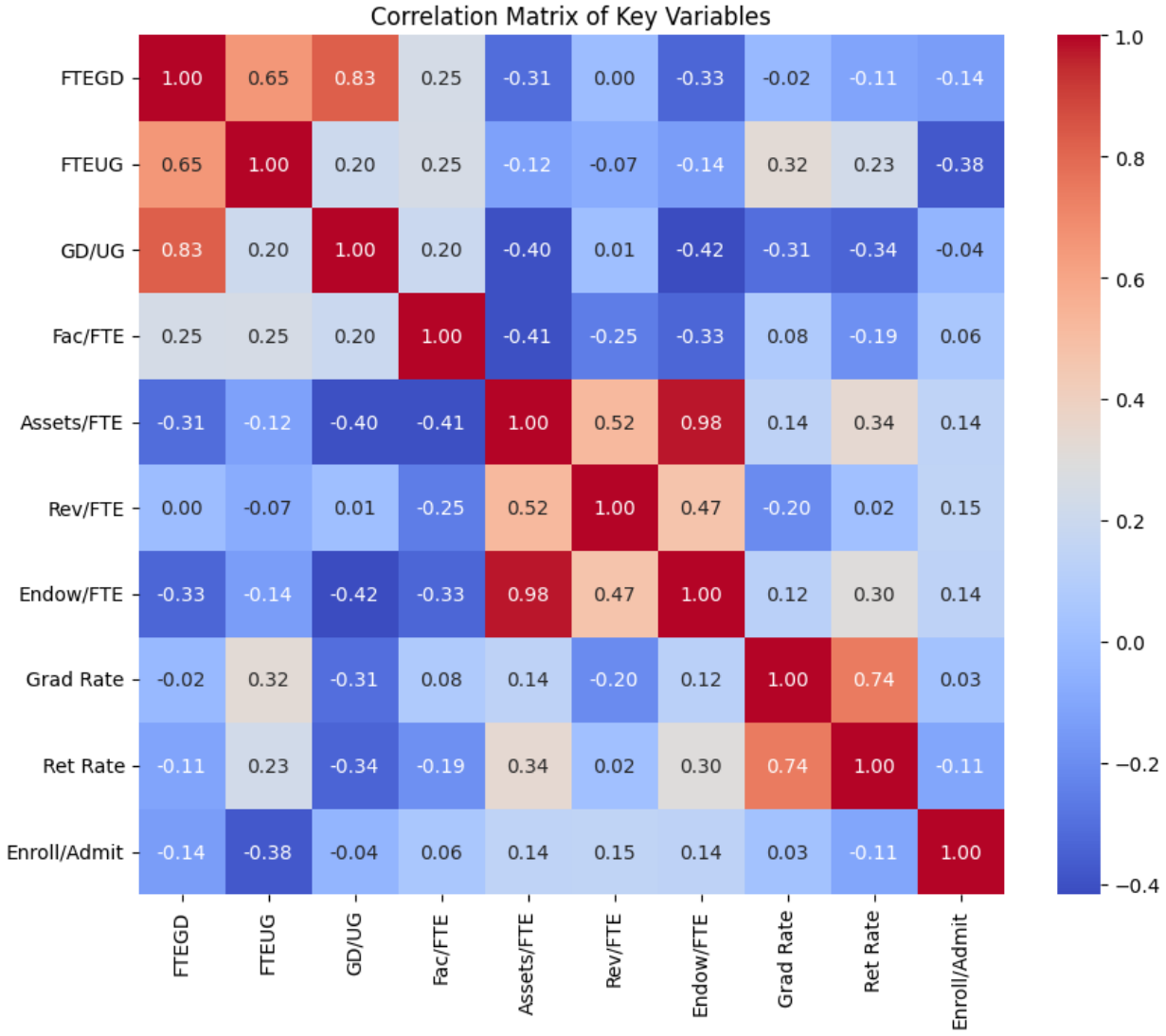
APPENDIX A

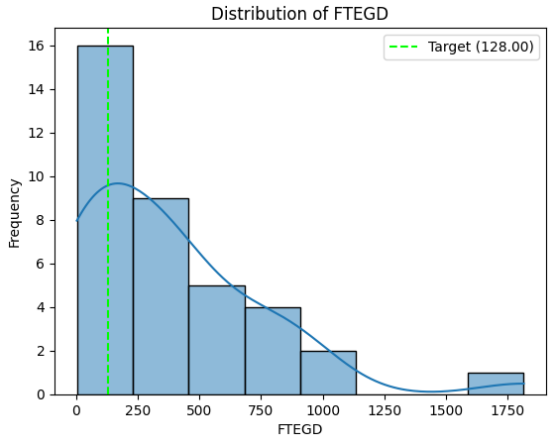


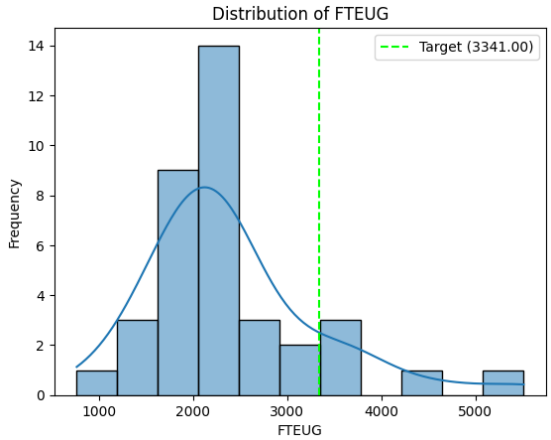
APPENDIX B

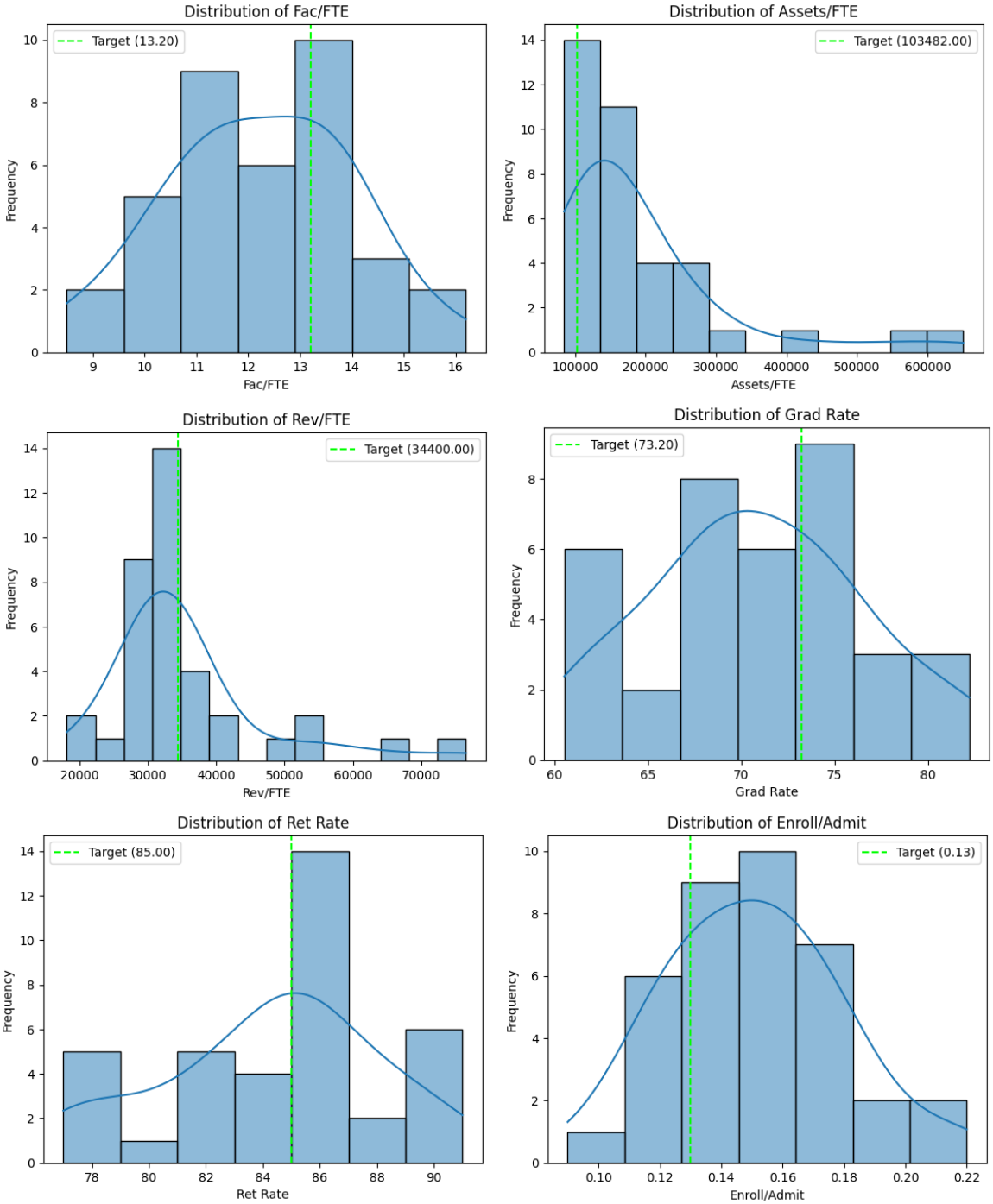

Distribution of distance between 2605 institutions and the target
Top 60 < 0.0183
Frequency
Gower Distance
0
50
100
150
200
250
300
350
400
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14

APPENDIX C


Correlation Matrix of Key Variables
FTEGD 1.00 0.65 0.83 0.25 -0.31 0.00 -0.33 -0.02 -0.11 -0.14
FTEUG 0.65 1.00 0.20 0.25 -0.12 -0.07 -0.14 0.32 0.23 -0.38
GD/UG 0.83 0.20 1.00 0.20 -0.40 0.01 -0.42 -0.31 -0.34 -0.04
Fac/FTE 0.25 0.25 0.20 1.00 -0.41 -0.25 -0.33 0.08 -0.19 0.06
Assets/FTE -0.31 -0.12 -0.40 -0.41 1.00 0.52 0.98 0.14 0.34 0.14
Rev/FTE 0.00 -0.07 0.01 -0.25 0.52 1.00 0.47 -0.20 0.02 0.15
Endow/FTE -0.33 -0.14 -0.42 -0.33 0.98 0.47 1.00 0.12 0.30 0.14
Grad Rate -0.02 0.32 -0.31 0.08 0.14 -0.20 0.12 1.00 0.74 0.03
Ret Rate -0.11 0.23 -0.34 -0.19 0.34 0.02 0.30 0.74 1.00 -0.11
Enroll/Admit -0.14 -0.38 -0.04 0.06 0.14 0.15 0.14 0.03 -0.11 1.00
FTEGD FTEUG GD/UG Fac/FTE Assets/FTE Rev/FTE Endow/FTE Grad Rate Ret Rate Enroll/Admit
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4


Distribution of FTEGD
Target (128.00)
Frequency
FTEGD
0 250 500 750 1000 1250 1500 1750
0 2 4 6 8 10 12 14 16


Distribution of FTEUG
Target (3341.00)
Frequency
FTEUG
1000 2000 3000 4000 5000
0 2 4 6 8 10 12 14


Distribution of Fac/FTE
Target (13.20)
Frequency
Fac/FTE
Distribution of Assets/FTE
Target (103482.00)
Frequency
Assets/FTE
Distribution of Rev/FTE
Target (34400.00)
Frequency
Rev/FTE
Distribution of Grad Rate
Target (73.20)
Frequency
Grad Rate
Distribution of Ret Rate
Target (85.00)
Frequency
Ret Rate
Distribution of Enroll/Admit
Target (0.13)
Frequency
Enroll/Admit