

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Volume 18, No.4
December 2025
ISSN: 1946-1836

In this issue:

- 4. User Experiences in a RAG-Empowered Application**
Shingo Kise, City University of Seattle
Sam Chung, City University of Seattle
- 14. Duality in 2D Apportioning: A Site Suitability Case Study for Spatial Data Analytics**
Peter Wu, Robert Morris University
- 20. Future Workforce Evolution - Impact of Artificial Intelligence Across Industries**
Nicholas Caporusso, Northern Kentucky University
My Hami Doan, Northern Kentucky University
Bikash Acharya, Northern Kentucky University
Priyanka Pandit, Northern Kentucky University
Sushani Shrestha, Northern Kentucky University
Rajani Khatri, Northern Kentucky University
Will Pond, Northern Kentucky University
Na Le, Northern Kentucky University
- 36. Affordable Housing in Florida: Systematic Literature Review and Exploratory County-Level Data Analysis**
Namratha Kulkarni, University of North Florida
Bharani Kothareddy, University of North Florida
Karthikeyan Umapathy, University of North Florida
- 46. Training a large language model to code qualitative research data: Results from discussions of ethical issues**
David Simmonds, Auburn University – Montgomery
Russell P. Haines, Appalachian State University
- 56. AI-Related Advertising on Facebook: Addressing Bias, Targeting Challenges and Regional Factors**
Sera Singha Roy, University of Melbourne
Tanya Linden, University of Melbourne
- 67. A Proposed Study of Factors Moderating Degree of Trust in LLM and ChatGPT-like Outputs**
William H. Money, The Citadel
Namporn Thanetsunthorn, The Citadel

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 three 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.

2025 ISCAP Board of Directors

Amy Connolly
James Madison University
President

Michael Smith
Georgia Institute of Technology
Vice President

Jeff Cummings
Univ of NC Wilmington
Past President

David Firth
University of Montana
Director

Mark Frydenberg
Bentley University
Director/Secretary

David Gomillion
Texas A&M University
Director

Leigh Mutchler
James Madison University
Director

RJ Podeschi
Millikin University
Director/Treasurer

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

Eric Breimer
Siena College
Director/2024 Conf Chair

Tom Janicki
Univ of NC Wilmington
Director/Meeting Planner

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

Copyright © 2025 by Information Systems and Computing Academic Professionals (ISCAP). Permission to make digital or hard copies of all or part of this journal for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial use. All copies must bear this notice and full citation. Permission from the Editor is required to post to servers, redistribute to lists, or utilize in a for-profit or commercial use. Permission requests should be sent to Scott Hunsinger, Editor, editor@jisara.org.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Editors

Scott Hunsinger
Senior Editor
Appalachian State University

Thomas Janicki
Publisher
University of North Carolina Wilmington

2025 JISARA Editorial Board

Queen Brooker
Metro State

Alan Peslak
Penn State University

Wendy Ceccucci
Quinnipiac University

Mark Pisano
Southern Connecticut University

Ulku Clark
Univ of North Carolina Wilmington

RJ Podeschi
Millikin University

Biswadip Ghosh
Metro State University

Asish Satpathy
Arizona State University

David Gomillion
Texas A&M University

Michael Smith
Georgia Tech University

Russell Haines
Appalachian State University

Katarzyna Toskin
Southern Connecticut University

Edgar Hassler
Appalachian State University

Karthikeyan Umapathy
University of North Florida

Melinda Korzaan
Middle Tennessee State University

Hayden Wimmer
Georgia Southern University

Li-Jen Lester
Sam Houston State University

Paul Witman
California Lutheran University

Muhammed Miah
Tennessee State University

David Woods
University of Miami Regionals

Stanley Mierzwa
Kean University

Daivd Yates
Bentley University

Juefei Yuan
Southeast Missouri State University

User Experiences in a RAG-Empowered Application

Shingo Kise
kisesHINGO1@cityuniversity.edu

Sam Chung
chungSAM@cityu.edu

City University of Seattle
Seattle, WA 98121

Abstract

With the widespread use of Large Language Models (LLMs) in various applications, there has been growing interest in leveraging their capabilities to improve user experiences and streamline processes. However, given the availability of various LLMs and Retrieval-Augmented Generation (RAG) systems, it is crucial to understand the differences between these technologies to effectively implement them and maximize their potential for providing better services. This paper investigates whether a RAG-empowered mobile app can enhance user experiences by delivering more relevant responses to specific user inquiries than a system equipped solely with an LLM. We focus on RAG, facilitated by LangChain, and compare its effectiveness to that of LLMs. As a demo case for comparison, we developed a chatbot-based internship placement system using React Native, integrating the ChatGPT API and LangChain for personalized, relevant responses. By evaluating chatbot responses using the RAG Assessment (Ragas) framework with metrics such as context precision, context recall, faithfulness, and answer relevancy to measure the quality of the RAG pipeline, we found that the RAG-empowered system consistently delivered more context-specific answers. A qualitative comparison revealed that the LLM system produced more generic responses than the RAG system. RAG systems can enhance the efficiency and effectiveness of internship placements by offering tailored assistance. Our findings highlight the potential of advanced NLP technologies to revolutionize applications such as chatbots, promoting innovation and enhancing user experiences.

Keywords: Large Language Model, LLM, Retrieval Augmented Generation, RAG, Case Study

Recommended Citation: Kise, S., Chung, S., (2025). User Experiences in a RAG-Empowered Application. *Journal of Information Systems Applied Research and Analytics*. v18, n4, pp 4-13. DOI# <https://doi.org/10.62273/WQTL4388>

User Experiences in a RAG-Empowered Application

Shingo Kise and Sam Chung

1. INTRODUCTION

Natural language processing (NLP) has revolutionized in recent years due to the development of large language models (LLMs), which provide remarkable capacities for text generation, understanding, and manipulation. As LLMs have become more popular, cutting-edge methods like Retrieval Augmented Generation (RAG) have also surfaced, which facilitate offering substitute answers for a range of NLP problems.

Although LLMs and RAG share the goal of solving complex linguistic problems, their underlying architectures, approaches, and performance results are quite different. LLMs generate responses based solely on pre-trained data, while RAG systems combine retrieval from external sources with generation, allowing for more accurate, specific, and up-to-date responses (Zhao et al., 2024). We challenge whether a RAG-empowered mobile app can provide better user experiences by delivering more relevant responses to specific user inquiries compared to one equipped with only an LLM.

As a demo case study, we choose an internship matching program. Students at a university can benefit from a system where they can gain valuable outside-of-class experiences, such as mentorship programs, internship matching programs, and more. Several case studies have been conducted on the development of web-based internship placement applications (Abdullah et al., 2017; Chaurasia, 2023). By applying an LLM or RAG to the system to guide each user based on their needs, we can significantly improve the quality of student's education experience and post-graduation job landing probabilities. This demonstration aims to showcase the potential of incorporating an LLM and an RAG into an internship placement system, instilling hope about the future of such systems.

The proposed approach entails developing a demonstration chatbot mobile application in React Native. This application seamlessly integrates with the OpenAI API to generate responses to user inquiries for the LLM system. Furthermore, LangChain is used within the application to provide answers to internship-related questions derived from files for the RAG

system. Users can engage in interactive and informative interactions by combining these advanced technologies and receiving relevant responses to their queries in real-time. This novel combination of an LLM and RAG aims to improve the user experience by giving users a thorough understanding of internship opportunities and facilitating smoother interactions with the internship placement system, making them feel more engaged and informed.

2. BACKGROUND

Recognizing the importance of internship experiences and seeking to improve accessibility, there has been some interest in using technology to refine internship placement processes and provide students with more valuable industry opportunities. Several studies have been conducted to develop web-based internship placement systems that are specific to each university. For instance, Chaurasia (2023) developed an application using Python, the Django framework, and a MySQL database to enable fast and easy access to placement procedures and related activities. Similarly, Abdullah et al. (2017) used Adobe Dreamweaver CC 2014 and **Cross-Platform, Apache, MariaDB, PHP, and Perl (XAMPP)** to address manual management inefficiencies in their internship program, which was plagued by cumbersome paper-based processes and insufficient coordination among academics, industry stakeholders, and students. In the same way, others developed a recommendation system using technologies like C#, JavaScript, CSS, HTML, and MySQL. While these studies have successfully developed applications and received positive usability feedback from their test users, it is important to note that the primary goal of these systems has been to simplify the internship program process rather than to increase internship acquisition rates or overall employment outcomes.

React Native is an open-source framework developed by Facebook for building cross-platform mobile applications. React Native, which uses the popular React JavaScript library, allows developers to create mobile apps using tools and concepts familiar to web developers. One of its primary benefits is the ability to write code once

and deploy it across multiple platforms, including iOS and Android, without compromising performance or user experience. React Native uses a declarative programming model, which allows developers to define UI components in JavaScript XML (JSX) syntax and then translate them into native User interface (UI) elements. This approach ensures that the resulting apps have a native appearance and feel, giving users a seamless experience (React Native RSS, n.d.).

LLMs are advanced AI models that have been trained on massive text datasets. These models, such as OpenAI's Generative Pre-trained Transformer (GPT) series, have transformed NLP tasks with their remarkable language understanding and generation capabilities. LLMs can comprehend and generate human-like text on a wide range of topics, making them extremely useful for a variety of NLP applications (Yang et al., 2024).

The OpenAI API, developed by OpenAI, gives developers access to cutting-edge LLMs, allowing them to use their powerful capabilities to build applications. The OpenAI API allows developers to integrate an LLM into their projects without having to train or fine-tune the models themselves. The API provides a variety of endpoints for tasks such as text generation, text completion, language translation, and more, allowing developers to easily incorporate advanced NLP capabilities into their applications (OpenAI platform, n.d.).

RAG is an AI framework that enhances LLMs with accurate, current information from an external knowledge base (Martineau, 2024). LangChain is a framework that simplifies the development of RAG applications by integrating LLMs with external data sources. It offers resources for building modular chains that incorporate data from databases and APIs. This approach enhances the relevance and accuracy of the generated content by grounding it in external knowledge (LangChain, n.d.).

Ragas is a framework used to quantitatively evaluate the performance of RAG pipelines. It provides metrics and methods to assess the effectiveness of RAG systems by comparing generated responses with ground truth answers (Ragas, n.d.).

3. RELATED WORK

The rapid advancement of LLMs has revolutionized the field of NLP, providing

unprecedented capabilities in addressing diverse NLP tasks and real-world applications (Yang et al., 2024). Integrating cutting-edge technologies such as OpenAI's GPT into real-world web and mobile applications has become more accessible in modern application development, providing practitioners with enhanced capabilities in NLP tasks (Odede & Frommholz, 2024). Furthermore, the development of RAG frameworks such as LangChain has made it easier to integrate LLMs into applications and helped improve LLM models. Recent studies by Gautam and Purwar (2024) have demonstrated that RAG systems can enhance the accuracy and relevance of responses, making them competitive with commercial LLMs. RAG enables LLMs to interact with various data sources and environments, allowing for model customization.

The application of LLMs and related frameworks in academia has led to practical innovations in the educational system. For instance, student assistant chatbots built on top of ChatGPT are a common application for LLM technology. These specialized chatbots, unlike traditional ones, focus on specific academic disciplines, providing tailored responses to student queries. The integration of OpenAI's advanced LLMs (GPT-3.5 turbo) with the LangChain framework and vector databases allows chatbots (*JayBot*) to provide detailed responses to questions about UK universities (Odede, 2024). Similarly, TA chatbots provide personalized support to computer science students, offering guidance and assistance without providing direct answers to their assignments (Liu & M'hiri, 2024).

4. METHODOLOGY

Our research will develop an RAG system using LangChain to improve the integration of LLMs with external data sources. We will also employ the Ragas framework, a novel tool for evaluating RAG systems, to assess our RAG pipeline. Furthermore, we will explore a unique application of RAG within an Internship Placement System, a context not previously explored in the literature. This approach aims to explore the capabilities of RAG implementation in academia to support students' success, contributing to academic and practical advancements in the field.

User Requirement

The approach begins by identifying the user requirements for the internship placement system. These requirements, including the need for seamless access to internship information, personalized assistance, and efficient navigation, were derived from a review of relevant journals

and literature, which focus on enhancing the process of obtaining internships for students (Hang et al., 2024; Menezes et al., 2022). Additionally, insights were gained from understanding the challenges faced by our classmates in securing internships. Understanding these requirements is critical for designing a system that caters to the needs of students seeking internships.

Design

The system architecture and design are developed based on user requirements to integrate LLM and RAG into the internship placement system. Both systems involve the development of a chatbot interface capable of interacting with users, understanding their inputs, and providing appropriate responses using the ChatGPT API. FastAPI is hosted on an EC2 instance on AWS with two endpoints, "gpt" and "intern." The "gpt" endpoint receives the user query, adds the prompt, sends it to the OpenAI server, and returns the response (Figure 1). The "intern" endpoint uses the LangChain framework, which adds an information retrieval component. Initially, it gathers data from a new data source using user input. Subsequently, the user's query, prompt, and relevant information are sent to the OpenAI server, and the response is returned (Figure 2).

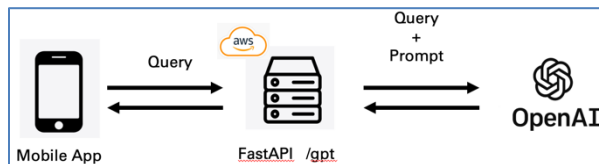


Figure 1: LLM Software Architecture

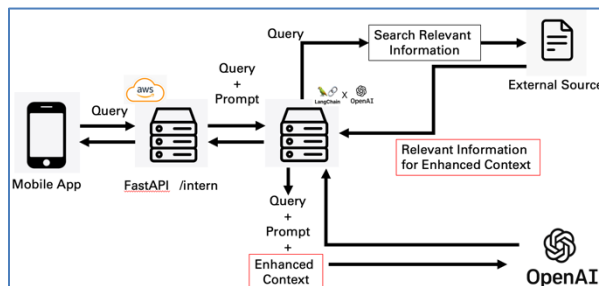


Figure 2: RAG Software Architecture

In addition, for containerization purposes, GitHub Codespaces was used, and for Continuous Integration and Continuous Development (CI/CD) environment, GitHub Actions were employed (Figure 3) (GitHub. n.d.) This automation tool facilitates various stages of the development workflow by enabling automated testing, building, and deployment processes. During the

development of the chatbot demo, this automation for DevOps significantly enhanced the overall efficiency and reliability of the development cycle by minimizing manual intervention and enabling faster iterations.

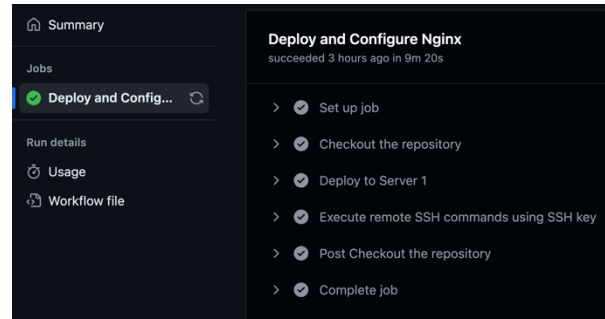


Figure 3: GitHub Actions for CI/CD Pipeline

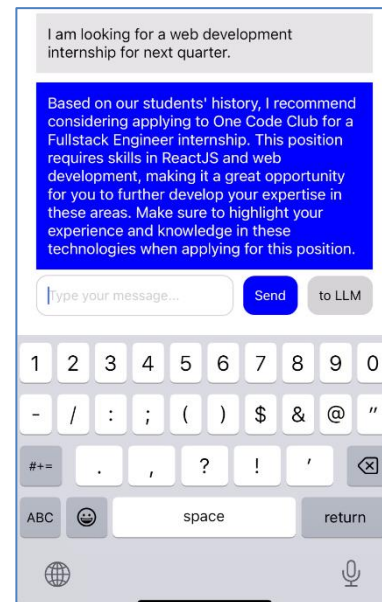


Figure 4: User Interface

Implementation

The internship placement system uses a variety of technologies to achieve its goals. The chatbot mobile application is developed using React Native (version 0.74.1), which provides a cross-platform solution for deployment on both iOS and Android devices (React Native RSS, n.d.) (Figure 4). Python (version 3.12.3) and FastAPI (version 0.105.0) serve as the backend framework, providing a robust and efficient environment for handling server-side logic, API creation, and database interactions. FastAPI supports non-blocking I/O operations, leveraging the asynchronous capabilities of Python's asyncio library to handle many requests concurrently and improve performance (FastAPI, n.d.). FastAPI is

hosted on an Ubuntu EC2 instance on AWS (Figure 5). The ChatGPT API (GPT-4) is the core component for NLP, allowing the chatbot to effectively understand and respond to user inquiries (OpenAI, n.d.). Furthermore, LangChain (version 0.1.10) supplements the system's functionality by retrieving internship-related information from files and databases, improving the overall user experience (LangChain, n.d.).

Name	Instance ID	Instance state	Instance type
chatbot-api	i-043235b5eac48a773	Running	t2.micro

Figure 5: EC2 Instance on AWS

5. DATA COLLECTION

Input Data

User Queries: Simulated user queries about CS internship opportunities are used. These queries are similar to the questions that CS students at a university might ask when seeking information on internship opportunities in their field. For example, "How do I apply for the internship course?" "I have completed a term at City University of Seattle. Am I eligible to apply for the internship course?"

```
intern.txt
1 Internship Eligibility:
2 For international students, after completing 3 quarters, you can
3 apply for the internship course with CPT as an elective course.
4 Students can take the internship course only if the school does
5 not offer two core courses, and if they have not already taken
6 two elective courses.
7 You need to obtain an offer letter and a program director's
8 approval letter by week 5 of the previous quarter.
9 You would not lose OPT time.
10 It is important to plan ahead and communicate with your advisor
11 and program director.
12
13 Past Students' Internship Record:
14
15 Student1:
16 Organization Name:
17 One Code Club
18 Industry:
19 Non-Profit Organization in Youth Education
20 Location:
21 Remote
```

Figure 6: A Sample Internship Description

Internship Descriptions: Sample internship descriptions or details are collected from various industries and fields, including information from past students' internship experiences, which are shown in Figure 6. These descriptions, tailored specifically for CS students, serve as input data for the LangChain component of the chatbot. The LangChain component plays a crucial role in processing these descriptions, enabling the chatbot to provide relevant responses to help students get internship opportunities. Importantly, the file accessible to the RAG system contains only general information about the university's CS students, ensuring that specific

student data remains protected and maintaining student privacy.

Response	Q. How do I apply for the internship course?
LLM	"To apply for the internship course, you typically need to follow these steps: 1. Visit the official website of the organization..."
RAG	"To apply for the internship course at City University of Seattle you need to obtain an offer letter and a program director's approval letter by week 5 of the previous quarter. It is important to plan ahead and communicate with your advisor and program director."
	Q. I have completed 1 term at City University of Seattle. Am I eligible to apply for the internship course?
LLM	"To determine your eligibility to take the internship course at City University of Seattle, it's best to refer to the specific requirements set by the university or your academic advisor..."
RAG	"For international students at City University of Seattle, you can apply for the internship course with CPT as an elective course after completing 3 quarters. It is important to plan ahead and communicate with your advisor and program director."
	Q. What's the name of the most recently completed internship by a student?
LLM	"I'm sorry, I am not able to provide real-time information about specific individuals or their activities. If you have any other questions or need assistance with general information, feel free to ask!"
RAG	"The most recently completed internship by a student was at One Code Club."

Table 1: Chatbot Response Comparison

Output Data

Chatbot Responses: The responses that the LLM system and the RAG system chatbot produced to the same questions from users. Table 1 presents the responses of the LLM and RAG systems to the same set of questions.

6. DATA ANALYSIS

We compare the specificity and relevance of the responses generated by each system. The LLM system's responses in Table 1 were not customized to the unique context of the students at the City University of Seattle; instead, they were generic. The general information in these responses can be applied to a wide range of institutions and contexts. On the other hand, responses from the RAG system in Table 1, which uses LangChain for information retrieval, were noticeably more precise. The RAG responses were customized to meet the unique requirements and circumstances of the City University of Seattle students.

To quantitatively evaluate the performance of these RAG pipelines, we employed the RAG Assessment (Ragas) framework, which helps assess the effectiveness of our RAG systems. While tools and frameworks exist for building these pipelines, quantifying their performance can be challenging. In order to ensure that the chatbot provides users with accurate and pertinent information, Ragas offers metrics and methods for assessing and quantifying the efficacy of our RAG pipelines

The evaluation process involves running predefined queries against the system to compare the generated responses with ground truth answers, calculating key metrics such as context precision, context recall, faithfulness, and answer relevancy using the framework (Ragas, n.d.). Table 2 in the Appendix shows the metric values for three questions asked to the chatbot demo.

Context Precision metric assesses how relevant the retrieved context is to the question, implying the effectiveness of the retrieval process. High context precision indicates that the context given to the model is highly pertinent to the question, guiding the generation of appropriate answers (Ragas, n.d.). In the application demo, the average context precision stood at 0.999.

The Context Recall metric assesses the retriever's ability to compile all the data required to respond to the query. It ensures the model can access all relevant information needed to respond thoroughly. A high context recall value indicates that the retrieval system effectively collects sufficient context for the given question (Ragas, n.d.). In the application demo, the average context recall stood at 1.0.

Faithfulness metric evaluates the degree of factual consistency of the answer with the context

related to the question. It ensures that responses from the chatbot are precise and appropriately matched to the provided context. High faithfulness denotes the answers' credibility and lack of errors or fabrications (Ragas, n.d.). In the application demo, the average faithfulness stood at 0.8889.

Answer Relevancy metric assesses the relevance of the response to the query. It ensures that the answers are accurate and directly relevant to the questions posed by the users. A high answer relevancy indicates that the responses effectively address the users' questions. In the application demo, the average answer relevancy stood at 0.9266.

7. FINDINGS

The data analysis revealed the chatbot's performance across all evaluated metrics, including Context Precision, Context Recall, Faithfulness, and Answer Relevancy. The RAG system, leveraging LangChain for retrieving external data, consistently provided more context-specific answers tailored to the specific needs of City University of Seattle users. In comparison, the LLM system generated more generic responses, which could be applied to a larger variety of users.

Throughout the assessment, these metrics, each representing a different aspect of the chatbot's functionality, repeatedly produced high numbers close to 1.0. High Context Precision indicates the model generated appropriate and accurate responses, as the retrieved context was highly relevant to the users' questions. Similarly, a high Context Recall value suggests that the retrieval system successfully gathered all data needed to offer comprehensive answers, ensuring comprehensive responses to user queries.

Furthermore, the chatbot's responses were consistent with the context given, as indicated by the high Faithfulness, which promotes confidence in the accuracy of the information provided. In addition, the high Answer Relevancy indicates that the chatbot regularly provided directly relevant answers to the users' queries. This performance can be associated with the chatbot's use of LangChain, which empowered the ability to understand and respond to user-specific queries precisely and accurately.

8. CONCLUSIONS

Overall, the availability of NLP-powered services, such as OpenAI API and LangChain, has provided

numerous opportunities for developers across multiple domains. The integration of LLMs and RAG into the internship placement system exemplifies this. It demonstrates the efficacy of using LLM and RAG to provide personalized chatbot services to students.

By evaluating key metrics such as Context Precision, Context Recall, Faithfulness, and Answer Relevancy, we have observed high levels of effectiveness in the RAG approach in providing accurate and relevant responses to user queries. Utilizing the power of LLM and RAG, the application can provide tailored responses to student inquiries, improving the user experience and facilitating seamless interactions within the system.

Our problem statement aimed to determine whether RAG can provide a better user experience compared to LLMs. These findings suggest that the adoption of cutting-edge technologies such as LLM and RAG has the potential to revolutionize NLP applications like chatbots. This would stimulate innovation and enhance user experiences for both individuals and enterprises.

9. VALIDATIONS

The validation of our research results and the chatbot demo performance experienced certain constraints due to the ten-week capstone course's limited duration. The time constraint significantly restricted our ability to develop and test the optimal RAG pipeline. Testing was conducted by simulating user inputs and evaluating the RAG pipeline using the RAGAS framework. However, accurate user testing would have provided valuable insights into user patterns and behavior, allowing for enhancements to the application and additional findings to improve the RAG system. Furthermore, the size and scope of our RAG input data were constrained. With a more extensive and diverse dataset, we could have performed more extensive and rigorous testing, allowing for a more comprehensive evaluation of the RAG pipeline.

10. FUTURE WORK

Future work should expand the RAG pipeline's training and testing dataset. By integrating a more extensive array of internship descriptions and user inquiries, it is possible to effectively verify the chatbot's capacity to provide accurate and relevant data. Further optimization and enhanced performance of the RAG pipeline will likely result from longer-term iterative testing

and refinement. Actual user tests must be included to obtain feedback and insights from real users. It will allow a more realistic evaluation of the chatbot's performance and user experience.

11. REFERENCES

- Abdullah, F. S., Abu Seman, A., Ibrahim, N., Majid, N. A., Abdul Wahab, N. M., Mustafa, M. S., & Mohd Sani, N. F. (2017). Web-based application of the internship management system. *Journal of Computing Research and Innovation (JCRINN)*, 2(3), 46-51.
- Chaurasia, S. (2023). Student Internship Placement Management System using python. *April-May 2023*, (33), 30-49. <https://doi.org/10.55529/ijrise.33.30.49>
- FASTAPI. FastAPI. (n.d.). <https://fastapi.tiangolo.com/>
- Gautam, B., & Purwar, A. (2024). Evaluating the efficacy of open-source LLMs in enterprise-specific RAG systems: A comparative study of performance and scalability. *arXiv*. <https://arxiv.org/abs/2406.11424>
- GitHub. (n.d.). Secure development made simple. <https://github.com/features/codespaces>
- Hang, K., Schrock, T., Ostrander, T. J., Berg, R., Menezes, T., & Wang, K. (2024). Industry mentoring and internship experiences at a Community College Baccalaureate Program in Software Development. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. <https://doi.org/10.1145/3626252.3630878>
- Langchain. LangChain. (n.d.). <https://www.langchain.com/>
- Liu, M., & M'Hiri, F. (2024). Beyond traditional teaching: Large language models as simulated teaching assistants in Computer Science. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. <https://doi.org/10.1145/3626252.3630789>
- Martineau, K. (2024, May 1). What is retrieval-augmented generation (rag)?. IBM Research. <https://research.ibm.com/blog/retrieval-augmented-generation-RAG>
- Menezes, T., Parra, A., & Jiang, M. (2022). Open-source internships with industry mentors. *Proceedings of the 27th ACM Conference on*
- Odede, J., & Frommholz, I. (2024a). JayBot -- Aiding University students and admission with an LLM-based chatbot. *Proceedings of the 2024 ACM SIGIR Conference on Human*

- Information Interaction and Retrieval*.
<https://doi.org/10.1145/3627508.3638293>
- OpenAI. (n.d.). <https://openai.com/>
- Ragas. (n.d.). Introduction.
<https://docs.ragas.io/en/stable/>
- React native. React Native RSS. (n.d.).
<https://reactnative.dev/>
- Yang, J., Jin, H., Tang, R., Han, X., Feng, Q.,
Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024).
Harnessing the power of LLMS in practice: A
survey on CHATGPT and beyond. *ACM
Transactions on Knowledge Discovery from
Data*. <https://doi.org/10.1145/3649506>
- Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y.,
Fu, F., ... & Cui, B. (2024). Retrieval-
Augmented Generation for AI-Generated
Content: A Survey. Retrieved May 20, 2024,
from <https://arxiv.org/pdf/2402.19473>

Appendix

Question 1	I have completed 1 term at City University of Seattle. Am I eligible to apply for the internship course?			
Answer	"Based on the provided context, for international students at City University of Seattle, you can apply for the internship course with CPT as an elective course after completing 3 quarters. It is important to plan ahead and communicate with your advisor and program director."			
Context	"Internship Eligibility: For international students, after completing 3 quarters, you can apply for the internship course with CPT as an elective course. Students can take the internship course only if the school does not offer two core courses, and if they have not already taken two elective courses. You need to obtain an offer letter and a program director's approval letter by week 5 of the previous quarter. You would not lose OPT time. It is important to plan ahead and communicate with your advisor and program director."			
Ground Truth	You are eligible to apply for the internship course after completing 3 quarters at City University of Seattle			
	context_precision	context_recall	faithfulness	answer_relevancy
	1.0	1.0	0.6667	0.9051
Question 2	How do I apply for the internship course?			
Answer	To apply for the internship course at City University of Seattle, you need to obtain an offer letter and a program director's approval letter by week 5 of the previous quarter. It is important to plan ahead and communicate with your advisor and program director.			
Context	"Internship Eligibility: For international students, after completing 3 quarters, you can apply for the internship course with CPT as an elective course. Students can take the internship course only if the school does not offer two core courses, and if they have not already taken two elective courses. You need to obtain an offer letter and a program director's approval letter by week 5 of the previous quarter. You would not lose OPT time. It is important to plan ahead and communicate with your advisor and program director."			
Ground Truth	You need to obtain an offer letter and a program director's approval letter by week 5 of the previous quarter.			
	context_precision	context_recall	faithfulness	answer_relevancy
	0.9999	1.0	1.0	0.8965
Question 3	What is the name of the most recently completed internship by a student?			
Answer	The most recently completed internship by a student was at One Code Club in the 2023 Spring Quarter, where the student worked as a Full Stack Engineer on frontend development of a website using Next.js.			

Context	“Internship Eligibility: For international students, after completing 3 quarters, you can apply for the internship course with CPT as an elective course. Students can take the internship course only if the school does not offer two core courses, and if they have not already taken two elective courses. You need to obtain an offer letter and a program director’s approval letter by week 5 of the previous quarter. You would not lose OPT time. It is important to plan ahead and communicate with your advisor and program director.”			
Ground Truth	One Code Club			
	context_precision	context_recall	faithfulness	answer_relevancy
	0.9999	1.0	1.0	0.9783

Table 2: RAG Assessment Values

Duality in 2D Apportioning: A Site Suitability Case Study for Spatial Data Analytics

Peter Y. Wu
wu@rmu.edu
Robert Morris University
Pittsburgh, PA 15108

Abstract

A common application in spatial data analytics is site selection. In site suitability study, we often construct circular or ring buffers around the site location to analyze the vicinity. To determine its suitability, we assess the impact of the feature attributes around the site location by apportioning the parts of every feature included in the buffer. In our case study, we are commissioned to find a grocery store site with sufficient resident population in its vicinity. To facilitate for our search with test-and-verify strategy, the traditional approach will determine the proportion of population in the buffer around the potential site location. Instead, we exploit the duality of the census block polygon and aggregate the resident population of the census block to its centroid, turning the 2D polygons of census blocks into the points of their centroids. With the markers on the centroids symbolized by the resident population, the construct serves as a better visual guide for our search. By aggregating the resident population to the centroid, the approach facilitates for our apportioning method to come up with an estimate of the resident population around the selected store site, verifying its suitability more efficiently.

Keywords: Polygon Duality, Site Suitability, Spatial Analytics, Geographic Information System, GIS.

Recommended Citation: Wu, P.Y., (2025). Duality in 2D Apportioning: A Site Suitability Case Study for Spatial Data Analytics, *Journal of Information Systems Applied Research and Analytics*. v18, n4 pp 14-19. DOI# <https://doi.org/10.62273/CZUH9195>

Duality in 2D Apportioning: A Site Suitability Case Study for Spatial Data Analytics

Peter Y. Wu

1. INTRODUCTION

Maps are traditionally made for reference use. The geographic information system (GIS) has given rise to analytic mapping: maps are made for analysis to achieve certain application purposes. Spatial data analytics is the study of large volumes of spatially related data. The modern GIS is presently trending toward the direction (Reddy, 2018; Dougherty et al., 2024).

Site suitability study is a common application in spatial data analytics. To determine whether or not a site location is suitable for a certain purpose, we gather the relevant information around the site. The data gathered is necessarily distributed spatially. We then construct a buffered region around the site to analyze the impact of relevant attribute information there. The buffered region is the site location extended with a circular buffer, or sometimes multiple layers of ring buffers for analysis of the impact with respect to proximity (Misra & Sharma, 2015).

Since the relevant attribute data may not be distributed evenly, the data may be provided in aggregation by partitioning regions, such as a polygon map layer over the area. We construct the buffered region around the site location, casting it over the polygon map layer to analyze the impact. Data apportionment is then the assessment of the portion of each polygon within the buffered region impacting the suitability (LaGro, 2013). While we present the relevant information to provide a visual guide for site selection, the data apportioning method will allow us to test and verify the site suitability. Repeating test and verify will lead our search to a feasible site location, and therefore we need an efficient and effective apportioning method.

This paper introduces an apportioning method which exploits the 2D duality of polygons. We convert the polygon map into a point map of the polygon centroids. We then aggregate the attributes of the polygon feature to the centroid. In the absence of a way to exhaustively search for a suitable site, we use the point map of centroids to visually guide our search, and we use the centroids to facilitate our apportioning

assessment for the suitability determination.

The next section has a brief survey of literature discussing our understanding of the current state of the art. Section 3 presents our case study of site selection for a grocery store in the downtown area, requiring certain resident population in the vicinity. Sections 4 and 5 describe our data gathering and how we set up our constructs to perform the spatial analysis. Section 6 discusses the benefits and limitations of our method. Section 7 closes with the summary of the paper.

2. LITERATURE SURVEY

Before the GIS started map making using the computer, maps were mostly for reference only. It was often too costly to make maps for the purpose of analysis (Schmidt, 1983). The GIS has become a very efficient and cost-effective tool for analytic maps (Heyward et al., 2006). Such is the trend for the GIS in spatial data analytics (Kanade, 2022). Site suitability is the assessment of factors and qualities to determine the site location for a particular activity. Site suitability study is a common application in spatial data analytics, facilitated by the use of the GIS (Jain & Venkata, 2007). Depending on the nature of the site selection project, there are varieties of different methods.

Spatial Environmental Model

The method usually applies to an area. The area is mapped and similar regions are identified with additional maps made. The maps incorporate factors relevant to study and a model is built for the purpose of comparison (Hopkins, 1977). The method requires comprehensive knowledge of the regions included and becomes onerous. It is not used often now.

Values Suitability Analysis

Geographical consideration in site suitability often extends to many other aspects. The ecological condition surrounding the site can be a natural extension, but aesthetic preferences as well as other human values, though subjective, also need to be taken into account quantifying costs and benefits (Reed & Brown, 2007).

Ordinal Priority Approach

The area is mapped according to the quality of the land, such as slope, soil type, vegetation, climate. Each quality is then quantified by a rating, and a method assigns the weight of each quality to determine the suitability (Mendoza et al., 2003; Melczewski, 2006).

In our site suitability case study, we have only one simple criterion: resident population. The issue involved then is the method of apportioning population in the neighborhood surroundings.

Apportionment Methods

The study of apportionment methods in the past decades has focused mostly on the politics of parliamentary seats allocation based on population (Kohler & Zeh, 2012; Koslap & Wilson, 2021) but not on the spatial distribution of population, which is related to political districting. That on the one hand depends on the politics of constitutional re-districting regulations, but it may also be because of the difficulty of apportioning based on 2D polygon intersections. When we have to rely on a visual guide to search for a feasible site location, all the more we need an efficient way to assess suitability so that we can easily repeat the test and verify procedure. For our case study, we propose a method described in the following sections.

we may need to augment the GIS with plug-in tools, or scripting support. If search for a suitable site location is necessary, there is often no exhaustive method available and we have to resort to search by repeated test and verify for suitability. Then we will want to have some kind of visual guide in the cartographic presentation and an apportioning method to assess the suitability [Misra & Sharma 2015].

3. CASE STUDY: SITE SELECTION

In our case study, we are commissioned by the city government in a project to revitalize businesses in the downtown area. While providing tax incentive to a potential grocery store, the city government asks us to provide a feasible store site location that meets the requirement of more than 1500 resident population within walking distance. We proceed to take up the project. Assuming the average walking speed of 3 miles per hour, the 10 minutes round trip walk means a range of $\frac{1}{4}$ mile, or 1320 feet. To create the study area for potential feasible site locations, we have the downtown area extended with the $\frac{1}{4}$ mile buffer. That is the area for a feasible store location. The only factor we need to consider is the resident population within the buffered zone. Figure 1 in the following page illustrates the study

area, buffered around downtown, with the census blocks over the entire area.

4. DATA GATHERING AND VISUALIZATION

We gather the data of residential population from the U.S. census. The U.S. government holds census every 10 years. The resident population is counted in every census block. We use the study area to select the census blocks in and around the study area, with the information of resident population in every census block. The study area helps us to focus on the census blocks relevant to our project.

We can then color shade every census block by the resident population to visualize the population pattern. The color shaded map is called a choropleth map (Dent et al., 2009). Since there is no full proof to exhaustively iterate through all possible store locations, we want the choropleth map to be a visual guide for us in the search for a feasible store location. We may then test the location as feasible by verifying the residential population in the $\frac{1}{4}$ mile vicinity over 1500.

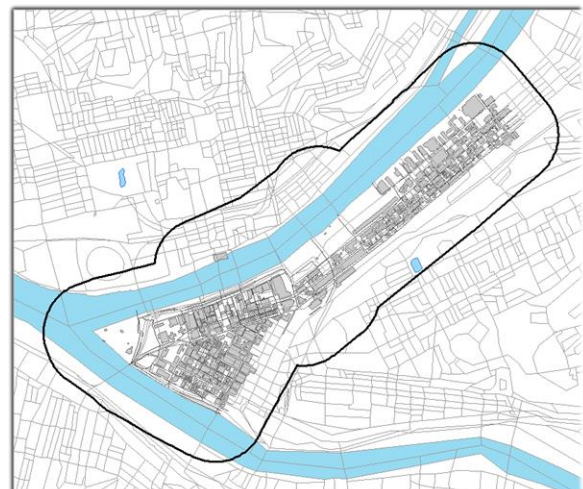


Figure 1: Study Area around Downtown

To estimate the resident population in the vicinity, we form a circular buffer region of $\frac{1}{4}$ mile radius at a potentially feasible store location. We need to prorate the apportionment of resident population in each census block that intersects with the circular region, assuming the population is distributed evenly within the census block and then we need to sum up all the portions of census blocks included. Figure 2 illustrates the study area with only the census blocks covering it, color shaded in the choropleth map by the resident population, and a $\frac{1}{4}$ mile radius circular graphic seeking a potential store location. The circular graphic requires us to calculate the prorated

areas inside the circle to estimate the resident population.

Since we need to search for a potential site to test and verify the resident population in its vicinity, our approach becomes too cumbersome to be conveniently effective. The presentation of the census blocks color shaded for population also does not serve as a very good visual guide.

5. SET UP FOR SPATIAL ANALYSIS

We change our strategy to exploit the 2D duality of our census blocks map by aggregating the entire population of the census block polygon to the centroid of the polygon. The centroid of a polygon is the center of gravity of the polygon. The following describes our approach.

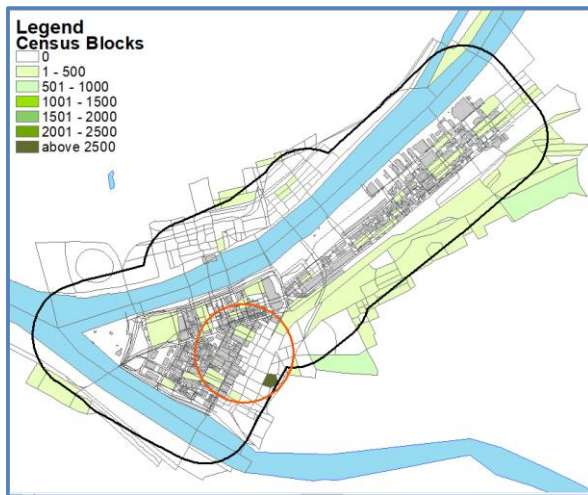


Figure 2: Census Blocks shaded into a choropleth map by population

We convert the polygon map of census blocks to a point map of the centroids. Since we can easily calculate the (X,Y) coordinates of the centroid of each census block polygon based on its geometry, we can form the point map of the census block centroids. Each centroid is then associated with the resident population of the entire census block. Instead of the choropleth map color shading the census blocks, we make a color marker at each centroid and code the color of the marker by the resident population of the census block. That provides a better visual guide to our search for a feasible store location than that of the choropleth map. Figure 3 illustrates the point map of the centroids using markers to show the resident population of each census block, and we can move around the circular graphic representing the buffer region in search of a feasible store site.

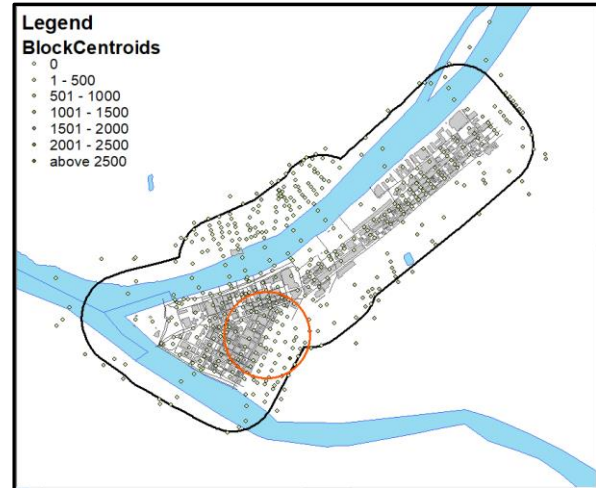


Figure 3: Census Block Centroids symbolized by population

Now to calculate the apportionment of the resident population in the vicinity of any selected location, we will consider only the block centroids which fall entirely within the 1/4 mile radius circular buffer of the potential store location. If the centroid falls inside the circular buffer, we include the entire population of the census block, and if the centroid does not fall inside, we disregard the entire census block. This approach is now illustrated in Figure 4: the centroids within the circular buffer are high-lighted, noting the selection of census blocks to be included in the apportionment of resident population in the vicinity.

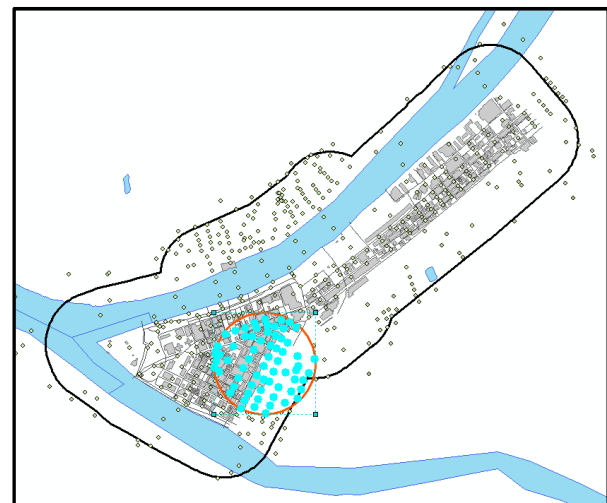


Figure 2. Census Block Centroids Selected.

From the attribute data table for the Centroids point map layer, since the centroids inside the buffer circle are selected, we can easily sum up

the Population now aggregated to the centroids. Figure 5 shows the attribute data table of the selected centroids with the population field highlighted for summation, to be done conveniently.

Granted that the randomly distributed pattern of resident population within the census block, the approach still calculates a reasonable estimate of the population in the vicinity, but it can be effectively done with convenience. We can repeat the procedure effectively and efficiently. The strategy facilitates for test-and-verify approach in the search for a feasible location, with an estimate of resident population in the vicinity.

OBJECTID	BlockID	POP2010	Shape_Length	Shape_Area	X	Y	Shape
51	0201001055	483	0.005065	0.000001	1341590.790267	410787.602278	Point
102	0201002021	303	0.009975	0.000006	1340795.818833	412070.907613	Point
41	0201001045	258	0.00504	0.000001	1342191.28324	410816.716118	Point
105	0201002024	157	0.004837	0.000001	1341859.88715	412506.911604	Point
50	0201001054	148	0.004803	0.000001	1342111.220796	410643.85196	Point
107	0201002026	95	0.003676	0.000001	1341977.113576	412156.593948	Point
14	0201001018	91	0.003075	0	1341773.967707	411861.26248	Point
36	0201001040	26	0.003967	0.000001	1340939.273541	410592.968477	Point
110	0201002029	6	0.001838	0	1341220.046116	411835.121351	Point
6	0201001010	0	0.003321	0.000001	1342652.837728	410846.807588	Point
7	0201001011	0	0.005289	0.000001	1342309.656389	411037.759895	Point
8	0201001012	0	0.005932	0.000001	1341814.507088	411184.742574	Point
9	0201001013	0	0.001778	0	1341678.885924	411389.637289	Point

Figure 3. Attribute Table with Centroids Selected

6. DISCUSSION

The benefit of our approach exploiting the duality of 2D polygons is that we do not have to go through the cumbersome process of calculating the prorated portions of areas of each census block polygon that crosses into the buffered region. By aggregating the total population of the census block to the centroid, we can effectively come up with an estimate of the resident population in the vicinity with ease. That allows us to search for a store site location with ease, performing test-and-verify strategy for a feasible site.

Our apportioning method turns the polygon map into a point map of polygon centroids. By aggregating the attribute information to the centroids, we can assess the apportionment of the attribute information to the selected site location with ease. The strategy may be generalized to apply to other cases when the attribute information is randomly distributed and aggregated in polygons in the area. Further research effort may attempt to formulate how to apply the methodology to other applicable situations.

However, our approach of aggregating the census block population to the centroid is different from prorating the population by the area of the census block inside the buffered region. We take into consideration only those census blocks that are within the buffered region. When the centroid is outside of the buffered region, the entire census block is not considered. We argue that the difference is acceptable for a reasonable estimate since the resident population is not necessarily evenly distributed within the census block, and is rather randomized. An analytical argument for this is due but there is substantial mathematical difficulty.

While we realize the difference in the apportioning methods, we consider the difference acceptable in practice. To make practical use of the ease of assessment supported by our test-and-verify search strategy, another approach of further research effort should therefore be directed toward calibrating the method in comparison to other traditional approaches. When we can compare the two different apportionment methods in many different cases, we may be able to formulate a scheme toward appropriate application of our approach as proposed.

7. SUMMARY

In the site suitability problem of spatial data analytics, we have to assess the impact of certain attribute information distributed over the area in the vicinity around the site location. Noting that the attribute information is not evenly distributed, it is aggregated to the polygons partitioning the area. We can construct a buffered region to define the boundary of its vicinity and then determine the apportioning of the attribute by proration of the area of each polygon inside the buffered region. The approach is however cumbersome to handle and does not effectively support test-and-verify search. In our case study, we are asked to seek for a grocery store location with the requirement of certain resident population in the vicinity. The resident population information is aggregated in census blocks in the area. Instead of calculating the proration by the area of each census block in the buffered region, we convert the census block polygons into a point map of centroids, attributing the population of each census block to its centroid. We can then just consider those centroids that fall within the buffered region to come up with an estimate of the population in the vicinity. The difference in the assessment seems acceptable, but the ease it allows offers us effective support for our test-and-verify search. The strategy may be generalized to similar cases of spatial data analytics.

8. REFERENCES

- Dent, B.D., Torguson, J.S., Holder, T.W. (2009) Cartography: Thematic Map Design (6th ed), McGraw-Hill.
- Dougherty, P., Burden A. and Blitz M. (2024). Human by Design: How AI Unleashes The Next Level Of Human Potential. Technology Vision 2024, in Accenture-Tech-Vision. <https://www.accenture.com/content/dam/accenture/final/accenture-com/document-2/Accenture-Tech-Vision-2024.pdf>
- Heywood, I., Cornelius S., Carver S. and Raju S. (2006). An Introduction to Geographical Information Systems, 2nd edition, Pearson.
- Jain, K. and Subbaiah Y.V., (2007). Site Suitability Analysis for Urban Development using GIS. *Journal of Applied Sciences* 7(18):2576-2583. <https://scialert.net/abstract/?doi=jas.2007.2576.2583>
- Hopkins, L.D. (1977) Methods for Generating Land Suitability Maps: A Comparative Evaluation. *Journal of the American Institute of Planners* 13(4):386-400. <https://doi.org/10.1080/01944367708977903>
- Kanade, V. (2022). What is Spatial Analysis? Definition, Working, and Examples. *Artificial Intelligence*, Spiceworks.com., August, 2022. <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-spatial-analysis/>
- Kohler, U. and Zeh J. (2012). Apportionment Methods. *The Stata Journal* 12(3):375-392. <https://doi.org/10.1177/1536867X1201200303>
- Koslap, K. and Wilson S. (2021). How Apportionment is Calculated. Published in *Census Blogs*, April 26, 2021. <https://www.census.gov/newsroom/blogs/random-samplings/2021/04/how-apportionment-is-calculated.html>
- LaGro, J.A. (2013) Site Analysis: Informing Context-Sensitive and Sustainable Site Planning and Design. John Wiley & Sons, Inc. Hoboken, NJ.
- Melczewski, J. (2006) GIS-based Multicriteria Decision Analysis: A Survey of the Literature. *Journal of Geographical Information Science* 20(7):703-726. <https://doi.org/10.1080/13658810600661508>
- Mendoza, G., Anderson, A. & Gertner, G. (2002) Integrating Multi-criteria Analysis and GIS for Land Condition Assessment: Evaluation and Restoration of Military Training. *Journal of Geographic Information & Decision Analysis* 6(1):1-16.
- Misra, S.K. & Sharma S. (2015). Site Suitability Analysis for Urban Development: A Review. *International Journal on Recent and Innovation Trends in Computing and Communication* 3(6):3647-3651 (ISSN:2321-8169). <https://ijritcc.org/index.php/ijritcc/article/view/4511>
- Reddy, G.P.O. (2018). Spatial Data Management, Analysis, and Modeling in GIS: Principles and Applications. In: Reddy, G., Singh, S. (eds) *Geospatial Technologies in Land Resources Mapping, Monitoring and Management*, vol.21. Springer, Cham. https://doi.org/10.1007/978-3-319-78711-4_7
- Reed, P. and Brown, G. (2007) Values Suitability Analysis: A Methodology for Identifying and Integrating Public Perceptions of Ecosystem Values in Forest Planning. *Journal of Environmental Planning and Management* 46(5):643-658. <https://doi.org/10.1080/0964056032000138418>
- Schmidt, A.H. (1983). Automated Cartography. *Computer Graphics*, T.L. Tunii (eds), Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-85962-5_30

From Angry Reviews to Classroom Success: Using LLMs to Synthesize RateMyProfessors.com Data

Nicholas Caporusso
caporusson1@nku.edu

My Hami Doan
doanm4@mymail.nku.edu

Bikash Acharya
acharyab2@mymail.nku.edu

Priyanka Pandit
panditp1@mymail.nku.edu

Sushant Shrestha
shresthas11@mymail.nku.edu

Rajani Khatri
khatrir2@mymail.nku.edu

Will Pond
pondw1@mymail.nku.edu

Na Le
len4@mymail.nku.edu

Northern Kentucky University
Highland Heights, KY 41099

ABSTRACT

In recent years, online professor review platforms have become increasingly prevalent in higher education. While previous studies have examined various aspects of these platforms, such as review sentiment and content validity, their potential as a source of information for academic success has been largely unexplored. This paper investigates the use of Large Language Models to analyze anonymous professor reviews and identify common themes related to effective teaching practices, course design, and student engagement. The goal is to provide students with actionable suggestions on how to succeed in specific courses rather than focusing on elements that do not directly impact educational outcomes. Our study analyzed reviews of nearly 40,000 computer science instructors, producing meaningful insights into course experiences. Although we realized our analysis of publicly available professor reviews the proposed methodology can be utilized in the context of official Student Evaluation of Teaching. We discuss how the proposed method can be utilized to process instructors' reviews, highlight teaching strategies, and elicit actionable information for both students and educators. Also, we describe how the same approach could also be utilized to identify areas for potential improvement.

Keywords: Large Language Models (LLMs), Education, Student Evaluation of Teaching, Educational Data Mining, Natural Language Processing, Mixed Methods Analysis.

Recommended Citation: Caporusso, N., Doan, M., Acharya, B., Pandit, P., Shrestha, S., Khatri, R., Pond, W., Le, N., (2025). From Angry Reviews to Classroom Success: Using LLMs to Synthesize RateMyProfessors.com Data. *Journal of Information Systems Applied Research and Analytics*. v18, n4, pp 20-35. DOI# <https://doi.org/10.62273/EBAP6519>.

From Angry Reviews to Classroom Success: Using LLMs to Synthesize RateMyProfessors.com Data

*Nicholas Caporusso, My Ham Doan, Bikash Acharya, Priyanka Pandit,
Sushant Shrestha, Rajani Khatri, Will Pond and Na Le*

1. INTRODUCTION

Education is continuously evolving, driven by advancements in technology as well as changing student interests, backgrounds, and learning preferences (Luxton-Reilly et al., 2018). It is important for instructors and institutions to understand teaching approaches and course design elements that resonate with today's learners to keep pace with these changes and provide an effective and engaging educational experience for students (Stephenson et al., 2018). Although there is a growing body of academic literature on pedagogical best practices, student voices and perspectives are often missing from this discourse (Robins et al., 2003). Indeed, student feedback is essential for professors to improve their teaching effectiveness and enhance learners' experiences. It helps professors identify their strengths and weaknesses, refine course design and content, adapt teaching methods and styles, address student concerns and challenges and promote student engagement and motivation, encourage self-reflection and professional growth, and align teaching with student needs and expectations.

To effectively solicit student feedback, institutions in higher education usually collect Student Evaluation of Teaching (SET) at the end of each course. SET are usually administered in the form of surveys with questions aimed at capturing students' view on aspects of teaching that are deemed as important such as clarity, competence, and classroom environment. Among various applications, SET provides instructors with insights that can be used to improve teaching quality and identify areas for professional development. In addition to feedback to professors, SET also informs personnel decisions at the administrative level (Coladarci & Kornfield, 2007). However, despite the widespread use of SET, the design of SET questionnaires and the analysis of the collected data often lacks a systematic approach, leading to fragmented and inconsistent utilization of the information across departments and institutions due to several factors. First, SET questionnaires generate a large amount of qualitative and quantitative data, making it challenging to

process and interpret the information effectively (Spooren et al., 2013). Specifically, qualitative data requires careful coding and analysis to identify common themes and patterns in student feedback. As a result, without a standardized approach to data analysis, different departments and institutions may employ varying methods, leading to a lack of comparability and consistency in how SET data is used (Uttl et al., 2017). Also, the results of SET evaluations are often not publicly shared, leading to a lack of transparency and consistency, and poor student involvement in the debriefing process. As a result, they are perceived as being primarily used for evaluating individual instructors' performance rather than identifying broader trends and best practices in teaching (Hornstein, 2017). This, in turn, limits the potential for SET data to inform institutional policies, professional development initiatives, and the sharing of effective teaching strategies across departments and institutions. Furthermore, this fragmented approach to SET analysis hinders the ability to derive meaningful insights and actionable recommendations for improving teaching effectiveness at a larger scale (Linse, 2017).

In the past decade, professor reviews platforms such as RateMyProfessors.com (RMP) have gained popularity because they address the unmet need of students to be able to access professor reviews before making enrollment decisions. Websites like RMP enable students to anonymously and publicly share their ratings, comments, and opinions on their teachers. To this end, and similarly to SET, RMP utilizes various quantitative criteria, including clarity, helpfulness, and easiness (Timmerman, 2008). Although its validity and usefulness have been questioned by scholars and educators, RMP offers a wealth of student reviews and opinions about courses and instructors. Indeed, RMP and similar platforms are not an official instrument, and SET surveys remain the most comprehensive and reliable source of student feedback for educators. Also, the reviews published on unofficial professor reviews websites are not moderated, and many reviews contain elements unrelated to pedagogy, including personal retaliation, inappropriate comments, and swear words. However, thanks to

their extensive publicly available longitudinal datasets, unofficial platforms like RMP could be utilized as a resource for experimenting novel solutions, particularly when SET data are not readily available, which is mostly the case. By examining RMP reviews, researchers focusing on SET can design, develop, and test novel systems for supporting students' and professors' experiences, ultimately potentially enhancing the quality of their instruction. Specifically, the similarities between SET and the reviews published on RMP make it possible to use RMP as a testbed to evaluate, for instance, whether solutions based on Natural Language Processing (NLP) can process unstructured information from textual and distill overarching themes and evidence-based insights.

This paper proposes a novel approach to analyzing data collected using SET surveys and extracting relevant information that can make it easier for professors, students, and administrators to draw insight from reviews. Specifically, in our work, we use Large Language Models (LLMs) and their capabilities in NLP tasks, including text classification and summarization. Our methodology utilizes an LLM-based pipeline that, starting from a large body of instructor reviews, (1) extracts a summary of the key dimensions and aspects of the learning experience (e.g., teaching style and classroom environment, learning approach and course content, participation and interaction, workload and expectations, and overall experience), (2) utilizes the key dimensions of teaching to process an instructor's reviews and generate a relevant summary of the learning experience provided to the students, (3) evaluates the students' sentiment on the dimensions of teaching effectiveness, and (4) suggests feedback for improvement. By leveraging the power of LLMs, our proposed method focuses on key pedagogical themes rather than on aspects that are not related to academic success, and it filters out irrelevant or biased information, including angry comments.

In our study, we focused on demonstrating the potential of LLMs and data-driven approaches to analyze a vast number of reviews, identify best practices, and offer practical guidance to students and professors. To this end, we validated our method using publicly available reviews posted on RMP. Nevertheless, the same approach can be utilized on official SET data.

2. RELATED WORK

SET surveys have emerged as the primary tool for

assessing teaching effectiveness in higher education. However, the rise of online platforms like RateMyProfessors.com has provided students with an alternative avenue to share their opinions and experiences with professors and courses. Although SET remains the most comprehensive and institutionally recognized source of student feedback, the lack of availability of data hinders researchers' ability to investigate students' ratings, comments, and reviews. Several studies suggested that universities should consider making their own SET data publicly available online to provide students with more representative and comprehensive data (Coladarci & Kornfield, 2007).

As a result, in the past decades, unofficial professor review platforms like RMP achieved increasing popularity thanks to their accessibility to end-users (i.e., students and professors) and researchers. To this date, RMP remains the largest dataset of professors' reviews, and it has attracted the attention of researchers interested in understanding its validity and potential utility for a variety of purposes.

In particular, several studies have explored the use of RMP data to gain insights into various aspects of higher education, overcoming the limitations of SETs in terms of public availability. Researchers have investigated the correlations between RMP ratings and traditional SETs (Coladarci & Kornfield, 2007), finding generally strong correlations, suggesting some degree of the validity of publicly available reviews as an indicator of instructor performance. Simultaneously, (Coladarci & Kornfield, 2007) found that RMP may be useful for identifying very highly rated instructors but less effective for differentiating among instructors with lower ratings and, therefore, that RMP is not a substitute for formal in-class evaluations. Other studies noted that easiness and quality ratings on RMP were positively correlated, suggesting that students tend to rate professors more favorably when they perceive the course as less challenging (Kindred & Mohammed, 2005). Several research groups conducted thematic content analyses of RMP comments and found that students often comment on both instructor competence and personal characteristics (Felton et al., 2008). Also, different studies (Kindred & Mohammed, 2005) analyzed the content of RMP reviews to identify common themes and factors that influence student ratings and found that students often mentioned professor personality, teaching style, and course difficulty as key factors in their evaluations, and they cautioned that RMP reviews should be interpreted with care, as they may not always reflect the actual quality of teaching. The

authors of a study (Legg & Wilson, 2012) found that students who voluntarily rate their professors on RMP tend to provide more negative evaluations compared to formal in-class evaluations. This self-selection bias raises questions about the representativeness of RMP ratings and their ability to reflect the overall student experience accurately. Also, other potential biases in RMP ratings have been a significant concern for researchers. Studies have shown that factors such as a professor's age, gender, ethnicity, and even physical attractiveness can influence student ratings on RMP (Legg & Wilson, 2012). The latter findings suggest the presence of biases and, consequently, raise questions about the fairness and objectivity of RMP evaluations and their impact on instructors' careers. For instance, (Gordon & Alam, 2021) found that students often comment on the accents of instructors with "Asian" last names, highlighting the potential for racial and linguistic biases in these evaluations. Additionally, some authors (Rosen, 2018) observed that professors in science, technology, engineering, and mathematics (STEM) fields tend to receive lower ratings on RMP compared to those in the humanities and arts, suggesting potential disciplinary biases.

Indeed, RMP has several limitations, and it should not be utilized as an official source of information for research regarding teaching experiences. However, its vast dataset of reviews offers an excellent resource for developing and testing systems before they are deployed on official SET. Despite the concerns regarding validity and biases that have been a subject of ongoing debate, RMP remains popular among students, with millions of users relying on it to inform their course selections (Boswell & Sohr-Preston, 2020). Also, it offers valuable insights into student perceptions and preferences. Also, in addition to students using RMP for enrollment decisions, instructors and institutions might already be leveraging the data available on RMP for hiring decisions or to gain insight on various aspects of teaching, including rapport with students, communication skills, and classroom management.

Another aspect that makes RMP's dataset an interesting source of information for research studies is the nature of the data collection process, which is not mediated by questions designed by a specific institution. Therefore, by analyzing RMP data, researchers can obtain a deeper and broader understanding of the factors that students consider important in their learning experience. This information can be used to

design solutions that improve teaching practices and enhance student satisfaction. To this end, although studies have suggested that RMP comments and qualitative feedback can provide insights into effective teaching practices (Hartman & Hunt, 2013), limited research has explored its use as a tool for identifying best practices in teaching. While the majority of research has centered on the validity and impact of RMP ratings, few studies utilized the content of RMP reviews as a source of insights for professors. One study utilized text analysis techniques to predict professor classifications based on student comments, revealing differences in the language used to describe "good" professors across various student groups. This study suggests that RMP reviews contain valuable information about student perceptions and priorities, which could be leveraged by professors to understand and adapt to their students' needs (Azab et al., 2016). The lack of studies analyzing the content of RMP reviews presents an opportunity for further research. By examining the themes, sentiments, and specific feedback contained within RMP comments, researchers could uncover actionable insights for professors looking to improve their teaching practices and better connect with their students. Such analyses could also shed light on the factors that contribute to student satisfaction and perceptions of teaching effectiveness, complementing the findings of traditional SET surveys.

More recently, AI techniques have been applied to analyze educational data and provide insights into teaching practices. The authors of a study (Sutoyo et al., 2020) used Machine Learning techniques, including sentiment analysis and natural language processing (NLP) frameworks such as BERT to analyze student comments from course evaluations. They identified key themes such as course content, teaching style, and assessment methods that influenced student satisfaction and learning outcomes. Their findings highlighted the importance of engaging students, providing clear explanations, and offering timely feedback. Also, the authors of (Wang et al., 2020) found that BERT was effective at identifying themes and sentiments in the comments, outperforming traditional machine learning approaches. These studies provided insights into student perceptions and learning outcomes in CS education and demonstrated the growing interest in using LLMs to analyze SET and RMP data. However, more research is needed to fully understand the potential and limitations of LLMs in this domain. Thus, there remains a gap in leveraging the rich qualitative data available in

RMP reviews to identify the best practices.

In this paper, we use the data from RMP as a testbed for an LLM-based solution ultimately aimed at processing reviews collected through SET surveys. Consequently, in our results we derive insights based on the content of the reviews from RMP to demonstrate the viability of our approach and validate our methodology rather than extracting information from the reviews. Nonetheless, the interaction dynamics of RMP, with specific regard to the ability of students to publish their comments anonymously, might also result in useful insights into learning experiences.

3. MATERIALS AND METHODS

The objective of our work is to automatically extract information from SET to enhance the assessment of professors' teaching quality to benefit instructors and students. Specifically, our goal is to leverage LLMs' capability of understanding and generating human-like text very accurately to analyze large volumes of unstructured data, such as student reviews of professors, whether from SET or other sources, and processing them in a way that provides instructors and students with more intuitive and actionable information.

In this paper, we present the results of a study in which we investigated the use of LLMs to analyze professors' reviews and extract key features that can inform and improve pedagogical practices as well as guide students in succeeding in academic courses. Instead of focusing on quantitative ratings such as professor quality, difficulty, and whether students would take the course again, our strategy takes a qualitative approach to the analysis of textual professor reviews, whether from RMP or official SETs. We designed a multi-step process for extracting different types of information from professor reviews, and we utilized publicly available data from an online website to validate our approach. To this end, the massive dataset offered by RMP is an exceptional testbed to evaluate different approaches based on LLMs, their feasibility, and their performances. In this phase, we are focusing on RMP because the nature of its data (i.e., the amount and it being publicly available) enables testing our method on a large number of reviews, validate our approach, and evaluate necessary improvements. In the next phase of our work, we aim to support data from official SET surveys. After defining key dimensions of teaching effectiveness and student success, our proposed methodology consists in using LLMs to process

individual professors' reviews, filter out irrelevant or inappropriate content, and extract the following outputs for each instructor or courses. The outputs are described in Figure 1.

1. A summary of the learning experience that students are expected to have with the professor or on the course and tips to perform well in the class. This primarily benefits students in their enrollment decisions, when they seek to know what kind of learning environment they will be in. In addition to providing prospective students with insights into course selections, this information can be utilized by the instructor to improve their teaching.
2. An analysis of the sentiment of the students, which can be utilized by professors to evaluate students' general perceptions and responses to their teaching style.
3. A list of actionable improvement items based on relevant students' suggestions. The instructor can use this information to quickly identify adjustments needed to accommodate an evolving audience.

In our study, we evaluated whether LLMs could assist in every step of this process, including summarizing a large number of reviews into an essential list of relevant feedback, capturing the expected classroom experience, and achieving insights that can be converted into suggestions for student success. By using LLMs, we aim to abstract aspects of the original reviews that can influence students and instructors negatively, such as the sentiment of the reviewer and their ability to articulate their opinions. Furthermore, this approach could also be utilized to filter out inappropriate information, including sexist comments (Boswell & Sohr-Preston, 2020), and provide the audience with a more polished digest.

In the context of official SET, each institution creates a survey with questions designed based on a predefined set of dimensions of teaching excellence and student success identified by a specific committee or unit. As a result, students' answers and reviews contain information collected from several questions each investigating one or more aspects. Therefore, using this top-down approach, the data collected from students' comments in the context of official SET reflect the aspects that are relevant for the institution. Conversely, in our case we utilized publicly available reviews collected in a bottom-up fashion from students. Therefore, the content was not guided or directed by any specific

dimensions, because RMP provides users with one text field only where they can enter their review. As we had no control over the data collection process, we could not make assumptions on the dimensions considered relevant by the students. Consequently, we used LLMs to also extract the most recurring topics and code and infer the relevant dimensions based on the content of students' reviews.

As a result, the steps in our process (also described in Figure 1) can be summarized as follows:

1. Collect the dataset. In the study presented in this paper, we utilized RMP's data. However, in regular application scenarios, the dataset is already collected by the institution and consists of course evaluations from SET.
2. Selection of professors. For the purpose of this study, we utilized a representative sample of RMP's dataset.
3. Pre-process the data to eliminate reviews that do not contain relevant information.
4. Extract the main themes from the content of the reviews.

Conversely, when applied to data from SET surveys, the process would be as follows.

1. Definition of quality metrics, which informs the creation of survey questions. Quality metrics would be defined top-down by the institution, whereas in our study they were extracted bottom-up from the content of the reviews.
2. Collection of the dataset, that is, administer course evaluations questionnaires to students and ensure a representative sample fills them out.
3. Process the data in a way similar to step 3 described above.

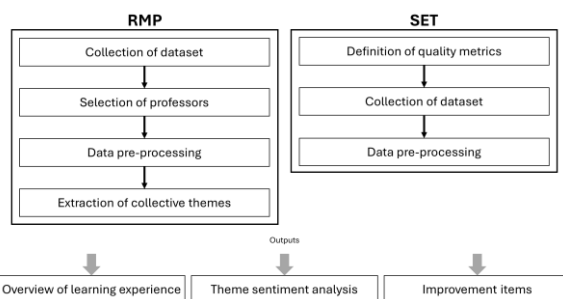


Figure 1 – An overview of our methodology and its different application with RMP's dataset and data from SET

3.1 Data collection

To obtain the dataset for our study, we developed software that automatically retrieved data from RMP using GraphQL, a query language for

Application Programming Interfaces (APIs). GraphQL enabled us to query RMP's server and specify the exact data fields required for our analysis. This approach allowed us to efficiently collect complete information about schools, professors, and their associated ratings. The initial dataset consisted of a total of 9,244 schools, 2,050,784 professors, and over 23,311,429 ratings.

After retrieving the initial dataset, we applied a filtering process to narrow the scope of our study to professors in one discipline only. We focused on a single academic field, that is, Computer Science (CS), to extract more targeted information and insights and actionable insights that are directly relevant to CS education. Therefore, we limited our dataset to 727,315 reviews from 227,687 individual CS courses taught by 49,147 professors at 3,502 schools. Nevertheless, the methodology could be utilized for other disciplines or generalized and applied in transdisciplinary fashion, regardless of a particular academic area.

Then, we aggregated and processed all the reviews on an individual professor basis. Although our initial goal is to process single courses, the data collected by RMP consists of very few reviews for most courses and in a large number of reviews in a limited subset of courses. As the high variance and sparse number of reviews per course would result in many courses having insufficient information, which would result in a poor outcome. However, this limitation would not affect data collected via SET, which has significantly higher response rates.

3.2 Pre-processing

Subsequently, we pre-processed our data to filter out irrelevant reviews. To this end, we analyzed the distribution of reviews per course and number of characters per reviews, which is shown in Figure 2. As shown in the image, a large number of courses have less than 3 reviews and less than 250 characters, resulting in very limited information. In fact, many students' comments involve just a few characters or a single word, or reviews such as "no comment", lacking useful information. Therefore, we removed a total of 12,099 professors whose reviews accounted for a total of less than 500 characters, regardless of the number of reviews, as shown in the first two lines of Figure 2. By doing this, we avoided analyzing reviews that, in addition to providing very little insight into the course experience, would cause the LLM to generate inaccurate content. Also, we removed a total of 2,471 professors with a large number of reviews

accounting for more than 12,000 characters in total. As these professors would take too long to process, we prioritized shorter reviews to test the feasibility of our system. Therefore, we restricted our initial analysis to a total of 34,577 professors (i.e., 70.35% of the dataset). As discussed earlier, we did not process individual course reviews because it would result in higher data sparsity in terms of the number of reviews and content and, consequently, limit the generalizability of our findings. In fact, reviews of 155,796 courses (i.e., 68.42% of the dataset) had less than 500 characters and, thus, would not be suitable for a comprehensive analysis.

This step was realized manually, by filtering reviews based on their length and content. Working with SET datasets would require the same type of preprocessing, which could be realized by analyzing the text with quantitative techniques or using traditional NLP approaches.

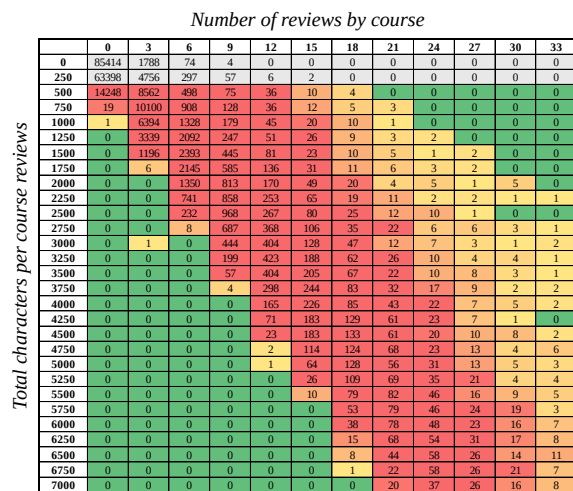


Figure 2 Distribution of reviews by number of reviews per professor and total characters (excerpt).

3.3 LLM selection

The third step in our process was to select an LLM suitable for text summarization, sentiment analysis, and text generation tasks. Many recent models, including free and open-source models, are equipped to perform well in these tasks. The goal of our work was to study the feasibility of our approach and validate our methodology rather than evaluating and comparing the LLMs performances of a series of models. As a result, our criteria in choosing the model were primarily guided by the feasibility integrating the LLM into the process. We decided to utilize Llama 3, an open-source LLM developed by Meta. Compared to its predecessors, Llama 3 exhibits better alignment with user instructions, leading to more

accurate and relevant responses, and offers a more diverse range of answers. Before choosing Llama 3, and specifically, the model trained with 8 billion parameters, we tested several other open-source LLMs, including Gemma, Mistral, and Phi3, on a subset of the dataset consisting of 100 reviews. Although their performances were similar, we chose Llama 3 because of its interoperability and openness to fine-tuning, which could be useful in our future work.

In our approach, we considered the LLM as a processing tool. Therefore, the model utilized in our study can be replaced by a different LLM that more appropriately or conveniently supports the specific use case or application scenario of the proposed approach.

In our study, we utilized the model on a client using Ollama, an open-source project designed to simplify the process of running LLMs on local machines. Ollama acts as a standard interface for interacting with an LLM, and it supports a growing number of models, many of which are Open Source. To process the dataset, we developed a custom JavaScript program that utilized Ollama's node package as an interface to query the LLM. The script was executed in a NodeJS environment on a computer equipped with a multi-core 12th gen Intel(R) i7-12800H processor with an NVidia RTX A2000 graphic card equipped with 8GB RAM and Cuda-enabled GPU.

3.4 Extraction of collective themes

In official SET surveys, students answer questions that investigate specific dimensions, which, in turn, can be utilized to guide the analysis of the content of the reviews. Conversely, as mentioned previously, one of the main limitations of using data from RMP as a test dataset is the unstructured way in which feedback is collected from users, with reviews being the result of one general text entry. In this context, prompting the LLM to analyze a review without any specific pointers results in a very general and inconsistent summary. Also, arbitrarily choosing dimensions of teaching excellence and student success would result in incorrect assumptions or in the LLM potentially generating text to fill out elements requested in the prompt that are missing in the data.

Therefore, we utilized the LLM to extract overarching themes that emerge across multiple reviews, in a process similar to manual coding in qualitative research. These themes could include common praise points, recurring concerns, or specific aspects of teaching that students frequently mention when providing feedback

about their professors. Identifying these collective themes helps understanding the broader patterns and trends. For example, themes could include the clarity of explanations, the helpfulness of feedback, the engaging nature of lectures, or the availability of resources.

To ensure the relevance and accuracy of the extracted themes, we initially extracted a set of pedagogical keywords and themes that guided the design of our system prompt to the LLM. To this end, we asked GPT-4 to analyze reviews for over 10,000 professors and extract key themes representing various aspects of teaching and learning. The LLM priming process involved an initial extraction of pedagogical keywords and themes from 10,000 rows of review data using GPT-4. This approach was validated through manual cross-verification to ensure that the themes accurately represented key dimensions of teaching quality, such as teaching style, student interaction, and assessment fairness. Figure 3 represents a word cloud of the most common elements found in reviews. This step was key to informing our coding process.

A number of themes emerged from the analysis of all the professors' reviews. We initially grouped them into 12 overarching areas related to teaching effectiveness and student success. These represent the key aspects that students frequently mention when providing feedback about their professors. While most themes are applicable across disciplines, industry, and real-world connections emerged as particularly relevant to CS education, especially in contexts such as software engineering.

1. Teaching methods and styles, representing whether the professor uses clear communication, structured learning, technology integration, interactive and hands-on activities, flipped classroom models, digital tools, multimedia resources, visual aids, animations, interactive lectures, and dynamic teaching techniques.
2. Course content and design, which incorporates real-world examples, updates content regularly, uses interdisciplinary perspectives, practical applications, case studies, varied assessments, project-based learning, reflective assignments, cross-departmental projects, current research, and podcasts.
3. Student engagement and participation, describing whether the instructor utilizes gamification, provides incentives, encourages active participation through discussions and coding sprints, and uses interactive simulations, real-time polls, collaborative learning, peer-to-peer teaching, peer review, student-led discussions, study groups, student showcase events, and infographics.
4. Feedback and assessment, which represents whether the professor provides timely and constructive feedback, uses clear grading rubrics, conducts formative and frequent assessments, offers self-paced learning options, sets transparent expectations, and monitors student progress.
5. Classroom environment and management, that is, whether the professor maintains a structured and respectful environment, uses inclusive teaching practices, creates an engaging atmosphere, maintains open communication, focuses on student-centered learning, uses active learning techniques, and adapts to different learning styles and paces.
6. Student support and development, including whether the instructor establishes mentorship programs, provides resources and support, promotes well-being, offers professional development, encourages growth mindset, continuously improves, fosters partnerships, provides growth opportunities, uses early alert systems, encourages learning from mistakes, and helps balance academic and personal life.
7. Collaboration and interaction, evaluating whether the professor assigns group projects, encourages collaboration, solicits student input, uses peer review, promotes collaborative projects, and uses communication platforms, online collaboration tools, and interactive workshops.
8. Use of technology in teaching, measuring how the instructor incorporates relevant technology tools and platforms, uses digital tools, integrates technology seamlessly, uses online learning platforms, virtual and augmented reality, learning management systems, and adaptive learning technology.
9. Content delivery and resources, describing teaching methods, the instructor's level of presence in the classroom, the use of modular assignments, online platforms, digital resource libraries, supplementary materials, recorded lectures, and optional workshops.
10. Industry and real-world connections, which are particularly relevant in technical disciplines, describing whether the instructor incorporates elements such as guest lectures,

builds industry connections, emphasizes real-world applications, aligns with professional standards, organizes guest speaker series, and collaborates with industry.

11. Continuous learning and improvement, representing whether students think that the instructor regularly updates content and methods, encourages professional development, promotes a growth mindset, implements feedback mechanisms, uses reflective assignments, and provides ongoing learning opportunities.
12. Flexibility and adaptability, representing whether the professor offers flexible deadlines, adapts teaching methods, uses adaptive learning technology, communicates expectations clearly, provides self-paced learning options, and implements early alert systems.

We did not quantify the occurrence of each theme in the reviews and weigh them based on the number of occurrences. This is because our goal was to identify all the key themes without necessarily setting a relevance threshold to scope the landscape of students' comments. Furthermore, associating any quantifiers to themes would introduce validity problems in our study, considering the concerns expressed by previous studies about the lack of completeness of RMP's data. Ultimately, this step was necessary only because of the characteristics of the RMP dataset.

Then, based on these pedagogical themes, we identified the following five dimensions that were most pertinent to a student's experience. This is to provide students with a more succinct summary highlighting the main aspects only.

1. *Teaching style and classroom environment*: the professor's teaching methods, ability to engage students, and create a conducive learning atmosphere define the classroom environment.
2. *Learning approach and course content*: the professor's organization and presentation of relevant, applicable course content, along with the use of assignments and projects, shape the learning approach.
3. *Participation and interaction*: whether the professor encourages student participation, being responsive to feedback, and maintaining availability outside of class characterize effective participation and interaction.
4. *Workload and expectations*: whether the professor establishes clear communication of course requirements, reasonable workload

distribution, appropriate academic challenge, and fair grading practices define the workload and expectations.

5. *Overall experience*: the overall classroom experience is determined by the professor's teaching effectiveness, ability to enhance student interest and engagement, supportiveness, and the sense of accomplishment students gain from the course.

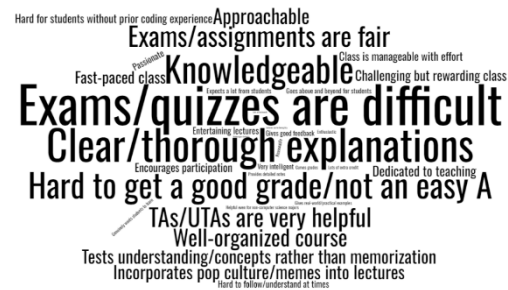


Figure 3 Word cloud of salient themes

3.5 Summary generation

After defining the five dimensions, we started feeding each professor's reviews into the LLM to generate a summary of their teaching experience. To this end, we used the five dimensions to generate the following system prompt, which was utilized to prime the LLM.

"You will be given a professor's review, and you will produce a description of the professor based on all the following aspects: - teaching style and classroom environment; - learning approach and course content; - participation and interaction; - workload and expectations; - overall experience. For each dimension, calculate a score from 1 to 5 based on the sentiment of the review. Absolutely describe all the 5 aspects. Finally, produce a list of suggestions for prospective students taking the professor, especially in computer science disciplines. Avoid mentioning the name of the professor and the reviews." This prompt was designed to elicit a comprehensive analysis of the professor's performance across five key dimensions, along with a numerical score for each aspect and a list of suggestions for improvement. The model was reset before processing each review to prevent any influence from previous inputs on the LLM's output.

3.5 Sentiment analysis

Subsequently, we analyzed the sentiment associated with each of the 13 initial themes, with the aim to determine whether the themes are generally addressed by students as positive, negative, or neutral.

Our goal was to evaluate whether the LLM is able to capture the overall sentiment of students with respect to each dimension to provide the instructor with a prioritized list of improvement items, if any. To obtain an analysis of the students' sentiment in each professor's class, we processed individual professors' reviews using the following system prompt.

"Based on the professor's review, rate each of the following aspects (if present) on a scale from -2 to +2 based on the sentiment of the review. 1. Teaching Methods and Styles; 2. Course Content and Design; 3. Student Engagement and Participation; 4. Feedback and Assessment; 5. Classroom Environment and Management; 6. Student Support and Development; 7. Collaboration and Interaction; 8. Use of Technology in Teaching; 9. Content Delivery and Resources; 10. Industry and Real-World Connections; 11. Ethics and Professionalism; 12. Continuous Learning and Improvement; 13. Flexibility and Adaptability. Only list the criteria and the numeric value".

4. RESULTS AND ANALYSIS

4.1 Summary generation

Processing all the reviews required approximately 84 hours. The LLM generated responses consisting of an average of 182.97 tokens, which took an average of 8805.30ms to generate.

The list below presents some examples of the summaries generated by the LLM based on the reviews. Specifically, the examples present excerpts of each of the five dimensions from different professors' reviews.

- *Teaching style and classroom environment: 5/5. The professor gives great lectures, uses YouTube videos as a teaching tool, provides easy-to-understand overviews of the weeks, and is a great teacher, which really helps you learn the material rather than memorize it.*
- *Learning approach and course content: 3/5. The professor is knowledgeable, but the way he teaches may require students to rely on outside sources or teach themselves the material. The course content is challenging in the second half.*
- *Participation and interaction: 3.5/5. While the professor is not overly engaging or interactive in his lectures, he does seem to be available for one-on-one office hours, which can be helpful for students who need additional support.*
- *Workload and expectations: 4/5. The professor sets realistic expectations for assignments and projects, with clear guidelines and deadlines. While the workload*

can be challenging at times, it is manageable, and students are encouraged to ask for help if needed.

- *Overall experience: 2.5/5. Students have mixed opinions about the professor's teaching style, with some finding him boring or dry and others appreciating his willingness to help.*

The examples show the LLM's ability to summarize effectively the key dimensions of professors' classroom experience, providing insight into aspects of teaching that can be useful for instructors and students. Simultaneously, the excerpts demonstrate the LLM's ability to abstract from reviewers' sentiments and produce a balanced reflection of teachers' qualities. For instance, in the fifth item of the list above, some students' reviews incorporated angry comments and inappropriate remarks, which the LLM successfully filtered out and reframed using a constructive tone. The latter aspect highlights the importance of using LLMs with high alignment and proper safeguards.

After processing the data, we assessed the LLM's output based on the following dimensions.

1. *Completeness*, that is, the presence of all the required elements, that is, (1) an analysis of each of the five key dimensions of teaching, (2) a numeric score for each dimension, and (3) the list of suggestions on how to succeed.
2. *Correctness*: whether the summary generated by the LLM reflected the content of students' original review.
3. *Consistency*: the LLM's ability to generate consistent output, including formatting of text, scores, and lists.
4. *Appropriateness*, including relevance of the information, use of an appropriate tone, and absence of inappropriate comments.
5. *Efficiency*, that is, the ability of the LLM to produce an effective summary without being too dry or verbose.

This post-processing step enabled us to evaluate the LLM's performance and, consequently, the feasibility and efficacy of our approach. To this end, using data produced from the post-processing parser described in the previous section, we analyzed quantitative dimensions (i.e., completeness, consistency, and efficiency) in all the 34,577 summaries generated by the LLM.

Figure 4 represents the completeness of the output of the LLM. Most summaries (i.e., 73%) included all five elements, whereas the remaining 27% lacked comments on one or more of the dimensions of teaching qualities. This is because

some students' reviews did not include comments that enabled the LLM to generate an appropriate summary. Also, 68% of LLM-generated reviews included a score for each dimension. A closer look at the content of some reviews revealed that although the information generated by the LLM is incomplete, in these circumstances, the system behaved correctly: instead of making up content, it simply avoided producing any. The score was completely missing in 19% of the reviews. This is because of the missing information described previously. However, in this case, the issue is also caused by an inconsistency in the results produced by the LLM. A mitigation strategy, in this case, would consist of either requiring the LLM to regenerate the review entirely or prompting the LLM to produce a score for each dimension present in the generated output. As far as the completeness of suggestions is concerned, the system provided two or more suggestions in 81% of the cases, whereas 17% of the reviews did not incorporate any recommendations. As in the previous case, this issue can be mitigated by requiring the LLM to process the original review and by deliberately asking it to only produce suggestions by conditioning the system prompt accordingly.

As far as the consistency of the output is concerned, our analysis primarily focused on syntactical aspects such as the formatting of lists and scores. LLMs produce Markdown-formatted output. Specifically, lists, including the dimensions of teaching quality and suggestions for academic success, were represented using the "-" symbol (i.e., unordered) and numbers (i.e., ordered) in 44% and 47% of the cases, respectively. In the remaining 9% of the cases, the output was unstructured. In the former situation, the parser was able to reconcile the items in the lists, in the latter scenario, the solution is to prompt the LLM to regenerate the output. Furthermore, when present (i.e., in 81% of the cases, as discussed above), scores were represented as a number (i.e., 3, or 5) in 42% of the cases and as a number with respect to its maximum value (i.e., 3/5, or 2.5/5) in 58% of the cases. The parser could handle such cases without requiring further processing.

For cases where the LLM-generated summaries were incomplete or inconsistent, a more detailed review revealed that this typically occurred in reviews with sparse content or ambiguous language. When a review lacked sufficient detail, the LLM occasionally omitted one or more dimensions of teaching quality, leading to incomplete summaries. Similarly, inconsistencies in formatting were more common in reviews with

non-standard phrasing or excessive repetition of themes. A potential strategy for improving incomplete output would involve prompting the LLM to regenerate the summary when key dimensions are missing. This could be achieved by setting minimum thresholds for data content, requiring the model to extract themes from multiple reviews rather than relying on sparse or brief input. Additionally, a fallback mechanism could request the LLM to provide suggestions for improving the reviews when a lack of data prevents a complete analysis, though this could result in content that is not present in the original review. Inconsistent formatting could be addressed through better prompt engineering. For example, by enforcing specific formatting rules within the system prompt (e.g., always use numbered lists for suggestions), we can ensure a more consistent structure across all outputs. Also, in our future work, we plan to integrate post-processing tools to standardize the final output format, resolving inconsistencies without requiring reprocessing of the original data. r professors with limited reviews, the LLM struggled to provide complete summaries due to a lack of data. One strategy to improve accuracy in these cases would be to aggregate reviews over multiple courses or time periods, allowing the LLM to analyze a broader dataset and generate more complete summaries. However, our strategy of choice is to include a fallback option to indicate that insufficient data is available to generate a fully detailed summary, ensuring that the output remains informative without misrepresenting the review data. We will implement this in our future work.

The last quantitative dimension considered in our analysis is the efficiency of the system, measured as the ability of the LLM to produce comprehensive reviews in a concise format. The average review length was 2054 ± 923 characters with a mode of 1926 characters. In 27,065 cases (i.e., 78%), the LLM generated reviews ranging between 1,000 and 3,000 characters, which is an appropriate length. In 3,394 cases (i.e., 9%), reviews were considered too short, whereas in 4118 instances (i.e., ~12%), they were too long.

Moreover, we evaluated correctness and appropriateness by sampling 500 LLM-generated reviews at random from six categories, that is, reviews with high and low completeness scores, consistency, and efficiency. As far as the correctness of the reviews is concerned, we did not find any LLM-generated summary that did not match the content of the original review. This is an indication of the performance of the LLM, its ability to limit hallucinations, and its high

alignment. Some items included in the suggestions consisted of general advice that was not necessarily part of the original review, which is not necessarily a concern, given the purpose of our approach. We found a strong correlation between appropriateness and the other dimensions of our analysis, with specific regard to completeness and consistency: out of the 500 summaries produced by the LLM and analyzed manually, all the outputs that scored 70% and above in the quantitative dimensions had appropriate content and did not raise any specific concern in terms of appropriateness. On the contrary, we found that in three cases, our LLM-generated summaries contained a somewhat negative tone resulting from the original student's comment, which was left unfiltered (e.g., "If you really wanna learn from the class, it's all up to you"). Based on our evaluation, these circumstances can be addressed by filtering out any output ranking low in completeness, correctness, and consistency.

	0	1	2	3	4	5
Summary	0.06	0.01	0.02	0.06	0.12	0.73
Score	0.19	0.01	0.01	0.05	0.06	0.68
Suggestions	0.17	0.02	0.09	0.28	0.25	0.19

Figure 4 Summary generation - Performance evaluation statistics

4.2 Sentiment analysis

Analyzing the sentiment of each professor's reviews took a total of 4 hours. The performance of the model was evaluated by randomly sampling approximately 10% of the output, that is, 3,000 professors, and manually comparing the content of the reviews and the extracted sentiment, assigning a score from 1 to 5 based on the accuracy of the LLM in classifying the sentiment. On average, the resulting score was 2.7, which was unexpectedly low, considering LLMs' ability in sentiment analysis tasks. However, the main issue was that in many cases, the model inferred a sentiment score for all the 12 dimensions even if the review did not have any content related to some of the teaching evaluation and student success metrics. This was due to the following factors:

1. The nature of the dataset and, specifically, the data collection process, which did not capture content for each of the dimensions. We extracted the 12 themes by aggregating the content of all the reviews. However, some themes were not mentioned in many professors' reviews. This issue is inherently solved using data from official SET surveys, where each dimension has a corresponding answer.

2. The number and specificity of the dimensions was too high for the LLM to find enough content in each review. This issue would be solved as in 1.
3. The inherent nature of LLMs, which makes them "fill in the blank" in case of missing or incomplete input. In addition to the solution mentioned in the previous two points, this issue might be solved by using prompt-engineering and fine-tuning techniques.

Overall, this aspect requires further investigation and will be analyzed in a follow-up study.

4.3 Improvement items

As the sentiment analysis could not provide an accurate representation of each professor's performance over the 12 dimensions, instead of analyzing improvement items on a per-professor basis, we aggregated the results and asked the LLM to analyze the sentiment on the entire set of aggregated reviews and identify suggestions for improvement. The results of our sentiment analysis (see Figure 5) show that, on average, students have an overall slightly positive attitude toward their instructors. Specifically, teaching methods and styles (+1.52) and ethics and professionalism (+1.50) received the highest positive sentiment scores. By manually comparing the reviews and the generated sentiment analysis score, we found that professors who employ engaging, interactive, and well-structured teaching methods while also emphasizing the importance of professional ethics are likely to be viewed favorably by students. Course content and design (+0.88), use of technology in teaching (+1.13), content delivery and resources (+1.13), and continuous learning and improvement (+1.06) also received positive sentiment scores, suggesting that students appreciate well-organized and relevant course content the effective integration of technology, accessible learning resources, and a commitment to ongoing improvement. Professors who keep their course content up-to-date, leverage technology to enhance learning experiences, provide comprehensive resources, and demonstrate a willingness to adapt and improve their teaching are likely to be positively perceived by students. Classroom environment and management (+0.78), student support and development (+0.63), collaboration and interaction (+0.69), industry and real-world connections (+0.69), and flexibility and adaptability (+0.50) received moderate positive sentiment scores. These results suggest that students value a positive and inclusive classroom atmosphere, supportive and development-oriented learning environments, opportunities for

collaboration, connections to real-world applications, and a degree of flexibility in the learning process. While professors are generally doing well in these areas, there may be room for further improvement to enhance student experiences and outcomes. On the other hand, student engagement and participation (-1.00) and feedback and assessment (-0.83) received negative sentiment scores, indicating potential areas of concern for students. These results suggest that students may feel less satisfied with the level of engagement and interaction in their courses and may desire more effective feedback and assessment practices. Professors should focus on strategies to promote active learning, encourage student participation, and provide timely, constructive, and actionable feedback to address these concerns and improve student sentiment in these areas.

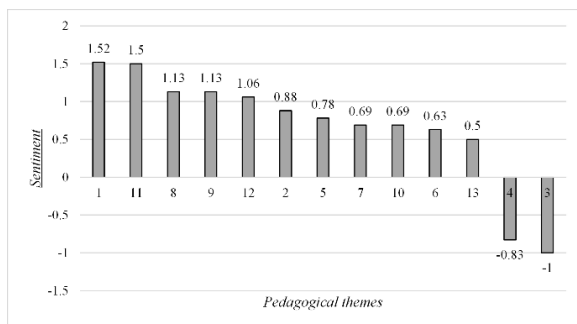


Figure 5 Sentiment analysis based on the aggregated sample of reviews

5. DISCUSSION

Our study focused on identifying key themes and aspects relevant to pedagogy in CS education, regardless of whether the reviews were positive or negative, by abstracting from arbitrary quantitative measures of teaching quality or bias caused by reviewers' sentiment. This approach has several advantages. By ignoring quantitative scores, the study provides a more comprehensive understanding of the key factors that influence student learning experiences. This holistic approach ensures that the identified themes are not biased towards only favorable aspects of teaching. Furthermore, considering both positive and negative reviews offers a balanced perspective on educators and their teaching practices. This approach acknowledges that even highly regarded professors may have areas where they can enhance their teaching, while professors with mixed reviews may still exhibit strengths in certain aspects of pedagogy. Finally, analyzing reviews across the spectrum of sentiment helps extract suggestions relevant to students'

academic success.

Indeed, our study suffers from the same limitations as other works based on RMP. As discussed in previous literature, publicly available reviews left spontaneously by a relatively limited number of individuals may not be representative of all experiences. For instance, students who are highly satisfied or dissatisfied may be more likely to leave reviews, leading to a potential bias in the data. Although this could influence the identified categories and themes captured in the paper and their relative importance, we addressed this concern by expanding our sample to many reviews across professors teaching different courses at numerous institutions. Furthermore, by abstracting from sentiment, our approach enables leveraging negative reviews as items students can consider. Another limitation lies in the LLM's ability to interpret subjective student feedback. While the model filters out inappropriate or biased language, there is still the potential for subtle biases in the data to influence the output. The LLM's reliance on sentiment analysis to score teaching dimensions may inadvertently overemphasize negative reviews, as students who are dissatisfied are more likely to leave detailed feedback.

It is important to clarify that the final dataset was indeed aggregated based on individual professors, but our objective was to distill general pedagogical themes rather than provide course-specific guidance. While this aggregation could limit granularity at the course level, we believe that patterns in teaching style, classroom engagement, and assessment methods often transcend specific courses. Thus, while the system produces summaries for professors across all courses they teach, these summaries reflect common pedagogical elements relevant to students' overall success. Nevertheless, we acknowledge this limitation and suggest future work could focus on extracting course-specific insights by refining the granularity of the data to individual course reviews, particularly for professors with a larger dataset of comments across various courses.

Another limitation in our study is related to the limited contextual information about the specific course, student background, or learning conditions. As the context is rarely captured in reviews, the lack of information could lead to an oversimplification of the complex dynamics of teaching and learning. Therefore, our analysis could fail to fully understand the factors contributing to a student's positive or negative experience. However, this problem is inherent in

other forms of evaluations of teaching, including SETs, which rarely capture contextual information. Nevertheless, the categories and themes identified in our study provide further studies with a taxonomy for qualitative and quantitative research studies on contextual factors, including courses, student demographics, and learning conditions.

Despite these limitations, the study's approach of focusing on key themes and aspects relevant to pedagogy, regardless of the sentiment of the reviews, provides valuable insights into the factors that shape student learning experiences in CS education. Educators can use these findings to reflect on their own pedagogical approaches and develop strategies to enhance student learning outcomes. Simultaneously, our approach provides prospective students with a more in-depth analysis of reviews left by past students, offering insight into the classroom experience and suggesting ways to prepare for the course. While previous studies analyzed RMP's reviews to investigate the dimensions of teaching, offering actionable items based on students' reviews is an original contribution to our approach.

Several aspects of our paper are innovative with respect to the state of the art. The previous use of RMP data has been limited to individual instructor evaluations without systematically identifying generalizable teaching themes across disciplines. Our approach differentiates itself by focusing on extracting broader pedagogical insights that are applicable across courses and instructors, aiming to provide actionable feedback to students on how to succeed in specific courses. This is in contrast to previous studies, which primarily assessed individual instructor performance based on RMP scores (Timmerman, 2008). By utilizing Large Language Models (LLMs), our methodology abstracts from the individual biases present in RMP reviews and identifies recurring pedagogical themes, such as teaching style and classroom management, which can inform both students and instructors.

Additionally, the literature demonstrates that RMP data can be biased by factors unrelated to teaching quality, such as professor attractiveness, gender, or discipline (Legg & Wilson, 2012). Our proposed method addresses these biases through a multi-step filtering process that removes irrelevant content, such as personal remarks or emotionally charged comments, ensuring that the focus remains on pedagogical aspects that contribute directly to educational outcomes. The LLM also abstracts sentiment and evaluates reviews based on

themes of teaching effectiveness, rather than subjective judgments that often dominate online evaluations.

While previous works, such as Sutoyo et al. (2020), have applied sentiment analysis and NLP frameworks like BERT to educational reviews, their focus was primarily on identifying sentiments and themes related to student satisfaction. Our study improves upon this by shifting the focus from student satisfaction to actionable pedagogical insights aimed at enhancing both teaching effectiveness and student success. Unlike sentiment analysis, which often overemphasizes emotional responses, our LLM-based approach seeks to provide a balanced and constructive analysis of teaching practices, offering not only a thematic breakdown but also concrete recommendations for both instructors and students. This methodological shift addresses the gaps left by prior studies, which often overlook the deeper pedagogical implications of student feedback.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented a study aimed at providing teachers and students with actionable insights into classroom experiences, to offer suggestions for improving the quality of teaching and, simultaneously, helping students succeed in their courses. To this end, we leveraged the vast amount of information available on RMP, a popular platform where students rate their professors on various criteria such as helpfulness, easiness, and quality of lectures. Several previous studies focused on the analysis of aspects such as the validity of the data collected by the platform, the assessment of professors' quality, and the sentiment of the reviews. On the contrary, our methodology introduces a novel approach to processing students' comments and extracting meaningful content that contributes to teaching effectiveness and student success rather than focusing on elements that do not directly impact educational outcomes.

To this end, after gathering the entire dataset of professor reviews, we filtered them to include only instructors teaching CS courses. Then, our analysis employed a mixed-methods approach based on the use of LLMs to analyze the qualitative reviews and the quantitative evaluation of the performance of the LLM. The primary objective of our study was to extract insights into teaching quality, professor-student interactions, and course content from user-generated reviews. We utilized large language models, particularly Llama3, for natural language

processing tasks to handle the vast amount of unstructured text data. Specifically, we asked the LLM to create a summary that represented the classroom through five key dimensions, that is, (1) teaching style and classroom environment, (2) learning approach and course content, (3) participation and interaction, (4) workload and expectations, and (5) overall experience. For each dimension, the LLM also assigned a quality score on a scale from 1 to 5 to provide students with a numeric indicator. Finally, based on the instructor's classroom experience, the LLM identified suggestions to help the students succeed.

Our findings demonstrate the potential of LLMs and data-driven approaches to analyze a vast number of reviews, identify best practices, and offer practical guidance for improving CS education and student outcomes. For educators, our analysis highlights effective teaching strategies and areas for improvement. For students, we offer suggestions and tips to excel in their chosen CS courses based on the collective experiences shared by their peers.

Based on the findings of this study, we propose several practical recommendations for implementing LLM-generated insights in educational practice. Educators could use LLM-generated insights as a complementary tool to improve their teaching practices. The summaries can provide a high-level view of student feedback, offering a more comprehensive understanding of their teaching effectiveness. The ability of LLM-based reviews to focus on recurring themes, such as classroom interaction and workload expectations, can help them make targeted adjustments that enhance student engagement and learning outcomes. As it relates to students, LLM-generated summaries can help students make more informed decisions when selecting courses or preparing for classes. By reviewing the pedagogical themes and recommendations, students can better understand what to expect in a course and how to succeed, rather than being influenced by the sentiment of the review, as reported by Boswell & Sohr-Preston (2020). For example, insights about workload expectations or participation requirements can help students plan their time more effectively. Finally, institutions could leverage LLM-generated insights to inform curriculum development and faculty evaluations. Thematic analysis of student feedback can identify broader trends in teaching quality, allowing departments to address systemic issues that may be hindering student success. Additionally, institutions could use these insights to develop professional development programs

tailored to the specific needs of educators, enhancing teaching practices across departments.

After validating the feasibility of our methodology, in our future work, we will apply our proposed method to data from official SET surveys. This would enable us to address some of the limitations we encountered using RMP's dataset, with specific regard to the lack of specificity with respect to key dimensions of teaching excellence and student success.

9. REFERENCES

- Azab, M., Mihalcea, R., & Abernethy, J. (2016). Analysing RateMyProfessors Evaluations Across Institutions, Disciplines, and Cultures: The Tell-Tale Signs of a Good Professor. *Social Informatics*, 438–453. https://doi.org/10.1007/978-3-319-47880-7_27
- Boswell, S. S., & Sohr-Preston, S. L. (2020). I checked the prof on ratemyprofessors: effect of anonymous, online student evaluations of professors on students' self-efficacy and expectations. *Social Psychology of Education*, 23(4), 943–961. <https://doi.org/10.1007/s11218-020-09566-y>
- Coladarci, T., & Kornfield, I. (2007). Ratemyprofessors.com versus formal in-class student evaluations of teaching. *Practical Assessment, Research & Evaluation*, 12(6), 1–15.
- Felton, J., Koper, P. T., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on ratemyprofessors. com. *Assessment & Evaluation in Higher Education*, 33(1), 45–61. <https://doi.org/10.1080/02602930601122803>
- Gordon, N., & Alam, O. (2021). The role of race and gender in teaching evaluation of computer science professors: A large scale analysis on ratemyprofessor data. *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 980–986. <https://doi.org/10.1145/3408877.3432369>
- Hartman, K. B., & Hunt, J. B. (2013a). What RateMyProfessors. com reveals about how and why students evaluate their professors:

- A glimpse into the student mind-set. *Marketing Education Review*, 23(2), 151–162. <https://doi.org/10.2753/MER1052-8008230204>
- Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, 4(1), 1304016. <https://doi.org/10.1080/2331186X.2017.1304016>
- Kindred, J., & Mohammed, S. N. (2005). "He will crush you like an academic ninja!": Exploring teacher ratings on ratemyprofessors. com. *Journal of Computer-Mediated Communication*, 10(3), JCMC10314. <https://doi.org/10.1111/j.1083-6101.2005.tb00257.x>
- Legg, A. M., & Wilson, J. H. (2012a). RateMyProfessors. com offers biased evaluations. *Assessment & Evaluation in Higher Education*, 37(1), 89–97. <https://doi.org/10.1080/02602938.2010.507299>
- Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, 54, 94–106. <https://doi.org/10.1016/j.stueduc.2016.12.004>
- Luxton-Reilly, A., Albluw, I., Becker, B. A., Giannakos, M., Kumar, A. N., Ott, L., ... Szabo, C. (2018). Introductory programming: A systematic literature review. *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 55–106. <https://doi.org/10.1145/3293881.3295779>
- Robins, A., Rountree, J., & Rountree, N. (2003). Learning and teaching programming: A review and discussion. *Computer Science Education*, 13(2), 137–172. <https://doi.org/10.1076/csed.13.2.137.14200>
- Rosen, A. S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors. com data. *Assessment & Evaluation in Higher Education*, 43(1), 31–44. <https://doi.org/10.1080/02602938.2016.1276155>
- Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598–642. <https://doi.org/10.3102/0034654313496870>
- Stephenson, C., Miller, A. D., Alvarado, C., Barker, L., Barr, V., Camp, T., ... Others. (2018). Retention in Computer Science Undergraduate Programs in the U.S.: Data Challenges and Promising Interventions. *ACM New York, NY, USA*. <https://doi.org/10.1145/3406772>
- Sutoyo, E., Almaarif, A., & Yanto, I. T. R. (2020). Sentiment analysis of student evaluations of teaching using deep learning approach. *The International Conference on Emerging Applications and Technologies for Industry 4.0*, 272–281. Springer. https://doi.org/10.1007/978-3-030-80216-5_20
- Timmerman, T. (2008). On the validity of Ratemyprofessors.com. *Journal of Education for Business*, 84(1), 55–61. <https://doi.org/10.3200/JOEB.84.1.55-61>
- Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54, 22–42. <https://doi.org/10.1016/j.stueduc.2016.08.007>
- Wang, W., Zhuang, H., Zhou, M., Liu, H., & Li, B. (2020). What makes a star teacher? A hierarchical BERT model for evaluating teacher's performance in online education. *arXiv Preprint arXiv:2012.01633*.

Affordable Housing in Florida: Systematic Literature Review and Exploratory County-Level Data Analysis

Namratha Kulkarni
n01380801@UofNorthFlorida.onmicrosoft.com

Bharani Kothareddy
bharani0223@gmail.com

Karthikeyan Umapathy
k.umapathy@unf.edu

University of North Florida
Jacksonville, FL 32224

Abstract

Despite the severe shortage of affordable housing in the United States, the social issue of homeownership is not well-understood and under-researched. We discuss two different studies in this paper. The first study presents the trends identified from articles relevant to affordable housing and data science approaches. The articles shortlisted for the first study were selected through a systematic literature review. The second study examines patterns and trends in housing characteristics, ownership, and occupancy at the Florida county level. More specifically, this research aims to understand the home affordability for renters vs owners using publicly available data. This paper analyzes housing data gathered from the U.S. Census Bureau. Using Tableau, we created interactive maps and visualizations to conduct exploratory data analysis and present the findings. Our analysis revealed significant disparities in housing conditions, ownership, and occupancy between different regions and counties in Florida. We also found that demographic variables such as age and income are key factors in relevance to housing patterns and trends.

Keywords: affordable housing, systematic review, exploratory data analysis, visual analytics, census datasets, demographic disparities.

Recommended Citation: Kulkarni, N., Kothareddy, B., Umapathy, K., (2025). Affordable Housing in Florida: Systematic Literature Review and Exploratory County-Level Data Analysis. *Journal of Information Systems Applied Research and Analytics*. v18, n4 pp 36-45. DOI# <https://doi.org/10.62273/JUTH9224>

Affordable Housing in Florida: Systematic Literature Review and Exploratory County-Level Data Analysis

Namratha Kulkarni, Bharani Kothareddy and Karthikeyan Umapathy

1. INTRODUCTION

Housing is one of the most important factors in the socioeconomic development of a thriving society. Effective housing supply is one of the government's policies to ensure effective affordability for people from all income levels, most importantly for the middle- and low-income levels (Adabre & Chan, 2019). Housing affordability refers to getting a particular standard of housing at a price or rent that does not impose an undue strain on household incomes in the eyes of a third party (typically the government) (McCrone & Stephens, 2017). Affordable housing as a social issue has been undergoing considerable changes in many countries for a number of years, with a steady reduction in the availability of housing units and associated support mechanisms (Czischke & van Bortel, 2023).

A number of terms have been used to explain housing forms that are affordable to middle- and low-income earners or income-constraint families. Depending on the country, several of these phrases could refer to various housing tenures. Affordable housing is a commonly used term in the United States and can refer to both rental and ownership housing. Public and social housing terms are more strongly associated with rental housing in the context of Europe. Cooperative housing is used in some European countries and refers to housing where the entire building is owned in common by a homeowners' association (Czischke & van Bortel, 2023). Despite the benefits of maintaining housing affordability and accessibility for socioeconomic growth, the global housing affordability dilemma continues to be unsolvable (Galster & Lee, 2021).

In both developed and developing countries, the housing unaffordability crisis continues to be a major challenge. In both rich and developing countries, the housing crisis is demonstrated by a lack of housing facilities, which has the consequence of promoting the construction of slums (overcrowded and dilapidated urban residential facilities with inadequate infrastructure) (Golubchikov & Badyina, 2012). Affordable housing has become a key issue impacting the lives of many residents, particularly

the population living in households that spend 40% or more of their disposable income on housing (Haffner & Hulse, 2021). Apart from the housing shortages, several of the existing affordable housing facilities are insufficient. If left unchecked, the housing scarcity and its inevitable outcome could get considerably worse, given that the world's population is expected to expand from 3.6 billion to 6.3 billion by 2050 (Golubchikov & Badyina, 2012). In response to the global housing problem, international organizations such as the World Bank and the United Nations (UN) have begun to develop policies to ensure enough housing. As a result, in addition to assuring housing affordability, achieving social and environmental sustainability goals for comprehensive sustainable development continues to be a hot topic in both developed and developing countries. The studies on affordable housing can broaden the knowledge base on which policymakers can base their decisions. More crucially, systematic and data-driven approaches to understanding affordable housing issues enhance the effect of existing knowledge on policy formation.

Research on affordable housing has mainly focused on descriptive explanations and policy causality analysis (Czischke & van Bortel, 2023). However, limited attention has been paid to understanding how affordable housing varies for population groups. To shed light on this gap, we conduct exploratory data analysis. As an exploratory data analysis would be effective when focusing on specific regions and for reasons for future work of the research team, we focus our research on Florida. Thus, this research paper aims to analyze the affordable housing situation in various counties in Florida using data from the U.S. Census Bureau datasets. The paper provides a detailed analysis of owner and renter-occupied housing units and the count of rooms in each unit for each Florida county. The population and demographic data of each county were also taken into consideration.

2. LITERATURE REVIEW OF DATA SCIENCE APPROACHES INVESTIGATING AFFORDABLE HOUSING

There has been a limited systematic review of the literature concerning data science and data analytics approaches to investigate the affordable housing context. Before we conducted our exploratory data analysis, we wanted to identify other literature that has created data visualizations and mapping of affordable housing trends.

Several systematic literature reviews on affordable housing have been conducted. Adabre and Chan (2019) conducted a review of 34 articles to determine the critical success criteria to attain sustainability in affordable housing. McCabe et al. (McCabe, Pojani, & van Groenou, 2018) focused their review on 67 studies in relevance to identifying the success factors, barriers, and motivations for the application of energy technologies in social housing. Anderson et al. (2003) reviewed 23 articles and reports that studied affordable housing and the reduction of segregation of low-income families. Although these reviews contribute greatly to the literature on affordable housing and other sustainability goals, they were not focused on articles that utilize data science approaches to address the affordable housing issue. Thus, we decided to conduct a systematic literature review to identify articles that investigate affordable housing issues using data science approaches. We aimed to gain knowledge on data science tools, methods, models, and techniques employed by researchers to study the affordable housing context. Thus, we learn from and make meaningful contributions to important social issues.

Systematic Literature Review: Data Collection

Google Scholar was chosen as the key scientific database for data collection. Keywords including "affordable housing" and "low income housing" were used in combination with "data analytics", "machine learning", and "data science" were used to ensure thorough data retrieval.

Publish or Perish software (Harzing, 2007) was used to get the details such as title, score, Article URL, abstract, cites per year, cites per author, author count, volume, issue, ECC, DOI, ISSN, Citation URL, Volume, Issue, Start page, End page, Cites per author, age, and Related URL.

Some limitation criteria were put in place to ensure the retrieval of adequate and manageable articles. The 'document type' was only limited to

journal articles because they form the broadest and collective knowledge base in any given academic field, given that we are seeking articles published by any researchers not just information systems researchers.

The search results were also refined to include articles published in the English language in the seven years ranging from 2015 to 2022, inclusive. We utilize the PRISMA flow diagram to depict the summarized view of the data collection and review article selection process. Figure 1 shows the PRISMA flow diagram. Aggregation of search results for different keyword combinations resulted in 911 articles.

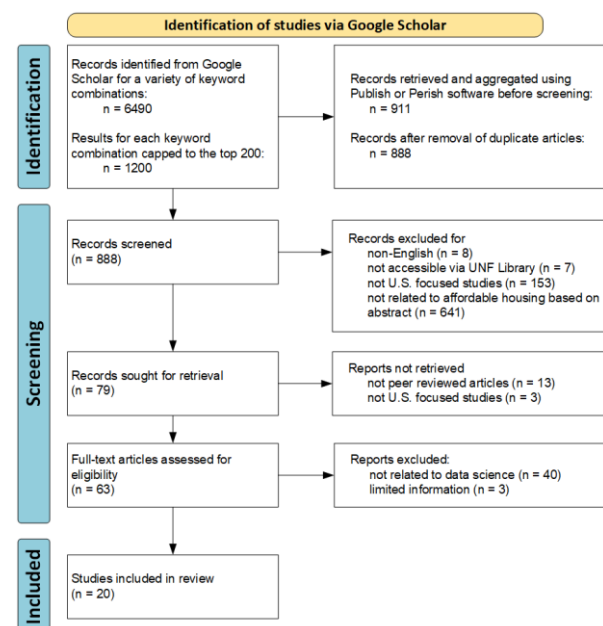


Figure 4 The PRISMA Flow Diagram of Search Result Screening

After the aggregation of search results, the duplicates based on the Title, Author, Abstract, and Article URL were removed, which resulted in 888 articles for further consideration. We then applied inclusion and exclusion criteria to filter the search result set further. The criteria we applied are 'If the article is English,' 'The article abstract is relevant to affordable housing,' 'Article is accessible through the university library,' and 'Article is related to US affordable housing.' After excluding articles that did not meet the criteria, we had 79 articles. We then excluded articles that were not peer-reviewed and not focused on the United States. We further assessed the remainder of the article by reading the title and abstract to ensure they investigated affordable housing

context using data science or data analytics approaches. After applying all of the exclusion criteria, the resulting articles are 20.

Systematic Literature Review: Data Collection

From the resulting 20 articles, we extracted the following details: data sources, tools used, analytical techniques applied, and research focus. After extracting relevant information, we created a categorical grouping of the information and calculated the frequency of the categories. Figure 2 depicts the frequency chart for data sources. Articles used multiple data sources in their research investigations. Federal and local government data sources were predominantly used by affordable housing researchers. In the figure 2, we show a breakdown of the sources for federal and local sources. It can be noted that Census, tax records, and Housing and Urban Development (HUD) data sources are key data sources for affordable housing investigations.

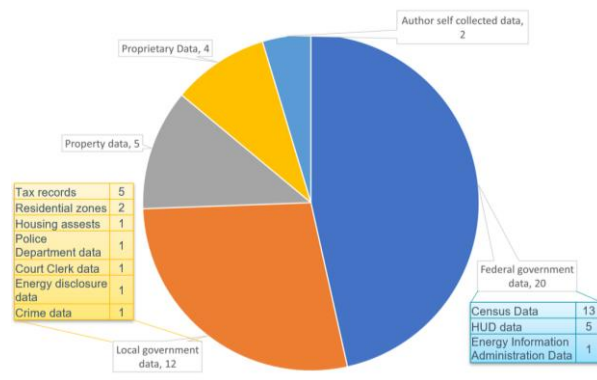


Figure 2 Data Sources

Figure 3 shows the frequency for data science tools specified in the articles. 50% of the researchers did not specify tools they utilized in their investigations. Authors, reviewers, and editors would have to ensure that tools used to perform analysis and solutions development are explicitly stated in the published articles; as these information are critical for replicability of the research. Of those specified, mapping tools which included GIS and other spatial tools, are mostly used by researchers.

Figure 4 provides counts of various data analytical approaches utilized in the identified articles. Some researchers did employ multiple techniques in their investigations. Our findings indicated that affordable housing researchers predominantly use regression analysis. Less usage of classification and clustering techniques indicates that researchers are using more statistical

analysis over machine learning techniques.

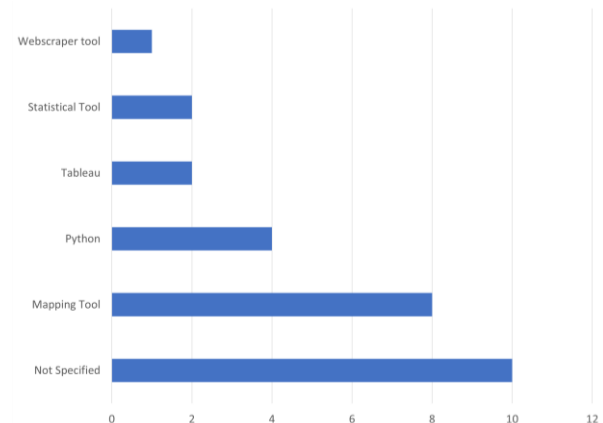


Figure 3 Data Science Tools

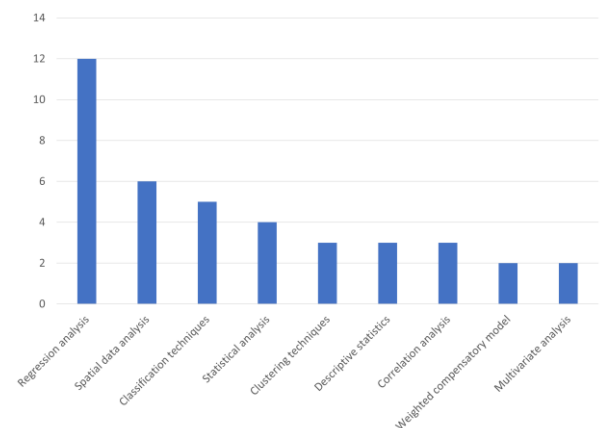


Figure 4 Data Analysis Techniques

With regard to key focus areas of the research, there were 12 articles where affordable housing was the primary focus, while for the remaining eight articles, affordable housing was the secondary focus area. In regards to data sizes, 11 articles did not specify the size of data used in their research, while five articles used big data and four used small datasets. Predominant of the articles (14 out of 20) did not partner with a community partner to study the affordable housing issue. The prevalent geographic focus for affordable housing research has been the city as the unit level of analysis, as shown in Figure 5. Despite the review of existing affordable housing literature conducted using Google Scholar and multiple relevant keywords, only a small portion of the result sets had articles that focused on affordable housing using data science approaches. Findings from extracted information reveal that there is minimal commonality among the published research articles. This implies that greater investment in affordable housing research

has a higher chance of making meaningful impacts for those affected by affordable housing issues.

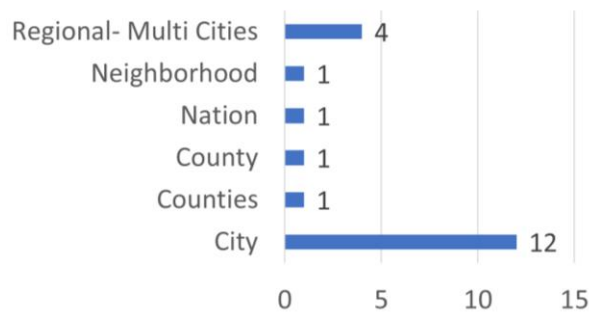


Figure 5 Geographic Focus

From the systematic literature, we noted there are limited studies focused on county as the unit level of analysis and none specifically focused on Florida. Thus, we are conducting an exploratory data analysis on affordable housing issues in Florida at the county level using publicly available datasets.

3. EXPLORATORY ANALYSIS OF AFFORDABLE HOUSING IN FLORIDA: DATA

Exploratory Data Analysis (EDA) is a crucial initial step in a data science study conducted by visualizing data to understand its characteristics, detect patterns & trends, and identify meaningful relationships. We use datasets from the Housing and Urban Development (HUD) and Census Bureau for the exploratory analysis.

As the literature has indicated, the housing cost burden and the ratio of owner-occupied to renter-occupied household units are closely associated with affordable housing issues. Thus, in this research, we aim to explore home affordability for renters vs. occupied household unit characteristics using publicly available data.

Datasets

In this section, we describe CHAS (Comprehensive Housing Affordability Strategy) and Census Bureau data profiles used in this research study.

Comprehensive Housing Affordability Strategy (CHAS)

The U.S. Department of Housing and Urban Development (HUD) obtains custom Census data from the Census Bureau that are not available through standard data releases. Using these data,

HUD produces the CHAS (Comprehensive Housing Affordability Strategy) dataset to reveal the extent of housing problems and needs for low-income households (HUD, 2023). The dataset, along with household counts, contains HUD Area Median Family Income (HAMFI), which is a median family income calculated by HUD to determine Fair Market Rents and income limits for affordable housing programs. Thus, CHAS data can be used to estimate the number of households that have affordable housing problems and have incomes low enough to qualify for HUD housing programs.

Census Bureau

As CHAS data was calculated based on the 2016 - 2020 American Community Survey (ACS) 5-year estimation data, we obtained relevant 2016 - 2020 ACS 5-Year Data Census Profiles for this research. We gathered Census data that are relevant to demographics, occupancy characteristics, housing units, and financial characteristics. We have listed the Census data profiles used for this research in Table 1.

Data Profile	Description
S2501: Occupancy Characteristics	Provides renters and owners occupied housing units based on household size, occupants per room, and age of householder
S2502: Demographic Characteristics for Occupied Housing Units	Provides renters and owners occupied housing units based on race, ethnicity, age, and education attainment of householder
S2503: Financial Characteristics	Provides renters and owners occupied housing units based on household income and monthly housing costs
S2504: Physical Housing Characteristics for Occupied Housing Units	Provides renters and owners occupied housing units based on housing unit structures

Table 1: Census Data Profiles used in the research study

4. VISUAL ANALYSIS OF AFFORDABLE HOUSING IN FLORIDA

Data described in the above section was downloaded for Florida counties. Downloaded data columns were merged based on the Florida

counties. In this section, we present mapping-based visuals for various data points gathered.

Population Distribution of Florida

We first created the population distribution of Florida as a baseline visual. Maimi-Dade is the most populous county, followed by Broward and Palm Beach counties. All three top populous counties in Florida are in the southeast region of Florida, stacked one over the other. The next three populous counties are Hillsborough, Orange, and Duval, which happen to be other metropolitan areas in Florida apart from Miami. The bottom five least populous counties are Liberty (being the least), Lafayette, Glades, Franklin, and Calhoun. The least populous counties are rural counties, which are mostly in the northwest and southwest regions of Florida. Figure 6 provides a shading map of population distribution for Florida counties.

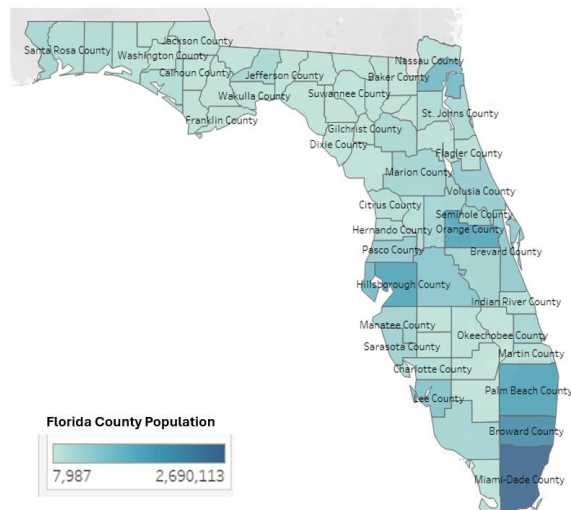


Figure 6 Florida County Population

Age Group Under 35

As the cost of home ownership and overall cost of living is increasing, younger population age group individuals would face several challenges to find affordable housing. Thus, we created a visual of the renter population for the age group under 35. We calculated a ratio of renters under 35 years old and the sum of renters of all age groups for each county. Figure 7 depicts the top 10 counties that have the highest renters ratio. Interestingly, we noted that Washington County has the highest ratio. Washington County is in the northwest region, and it is a rural county. The highest ratio is due to a high number of the younger population but less number of homeowners, leading to the majority of the younger population renting.

In a similar vein, we calculated the ratio of owners under 35 and the sum of owners for all age

groups. Figure 8 shows the top 10 counties that have the highest owners ratio. We note that Collier County, where Naples is located, is the highest, followed by Duval, where Jacksonville is located, then Hillsborough, where Tampa is located, and then Orange County, where Orlando is located. Thus, it seems younger populations tend to become homeowners in metropolitan cities like Jacksonville, Tampa, and Orlando more than high the populous Miami metropolitan area.

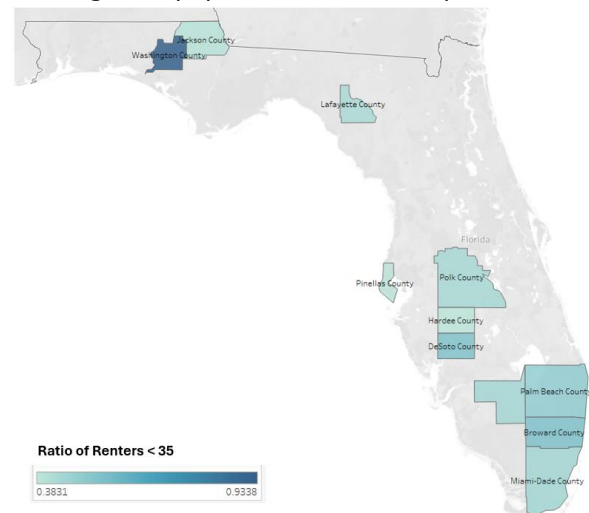


Figure 7 Top 10 Counties with Highest Renter Ratio for Age Group under 35

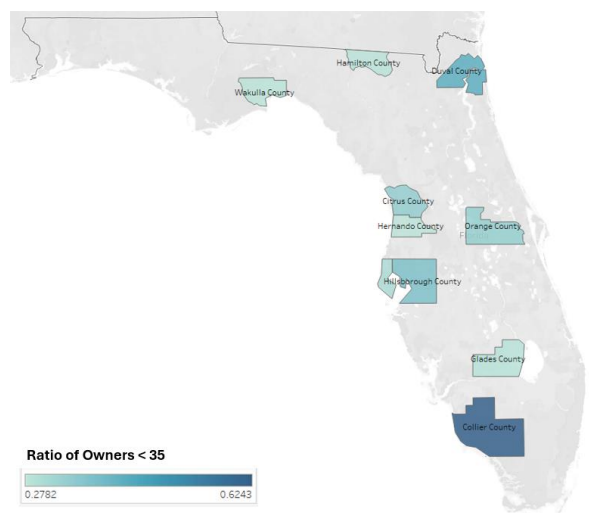


Figure 8 Top 10 Counties with Highest Ratio of Owners for Age Group under 35

Renters vs. Owners Home Built Year

Next, we created visuals based on the year the home was built. We focus on newly built home as it adds to the supply of available homes and would be preferred to live in due to lesser maintenance costs. We created a ratio of renters living in homes built in 2019 and after and all

renter homes, and similarly, we created a ratio of owners living in homes built in 2019 and after and all owner homes. Figures 9 and 10 show shading maps for Florida counties for the ratio of renters and owners, respectively, for homes built years after 2019. Regards to newly built homes with the highest renters ratio, Okaloosa, Leon, and Bradford counties in north Florida, along with Charlotte and Hernando counties, were in the top 5. All of the top 5 counties were rural counties. Regarding newly built homes with the highest owners ratio, Palm Beach County, which is north of Miami, has a large number of condos. Followed by Wakulla County, which is a rural county that has been seeing a growing population and growth of newly built homes.

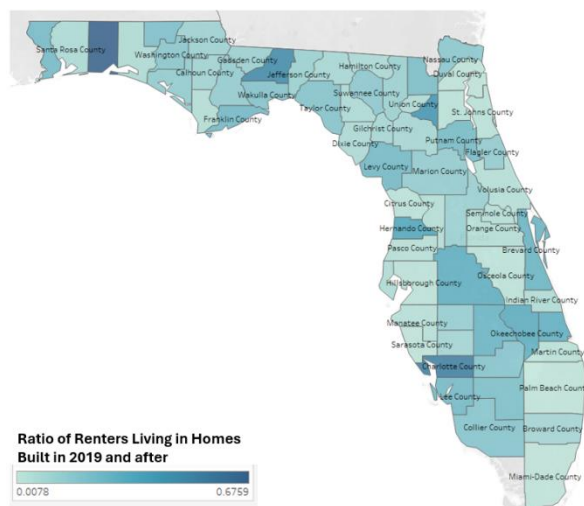


Figure 9 Renters Home Built in 2019 and after

Income Constrained Household

Median household income for Florida as Census ACS has been around 55K for the past few years, including 2020 (FLHealthCharts, 2024). Thus, we focus our attention on household income in the range of 25K to 50K. Exploratory analysis of the number of renters with incomes in the lower income buckets may indicate a need for affordable housing programs or rent control policies. We created a calculated field of ratio of owners with household income 25K to 50K and renters with household income 25K to 50K. From Figure 11, the majority of Florida counties are in light teal, indicating that the majority of income-constrained families are renting. Calhoun and Pinellas counties are the top two counties that have higher rates of those who are income-constrained who are homeowners. In contrast, in other counties majority of income-constrained households are renters

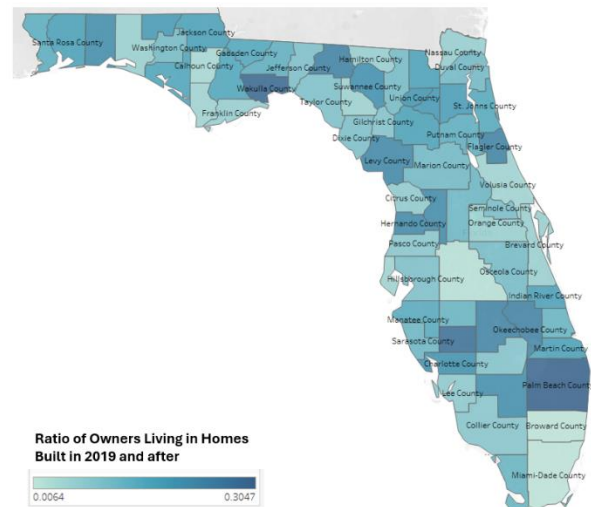


Figure 10 Owners Home Built in 2019 and after

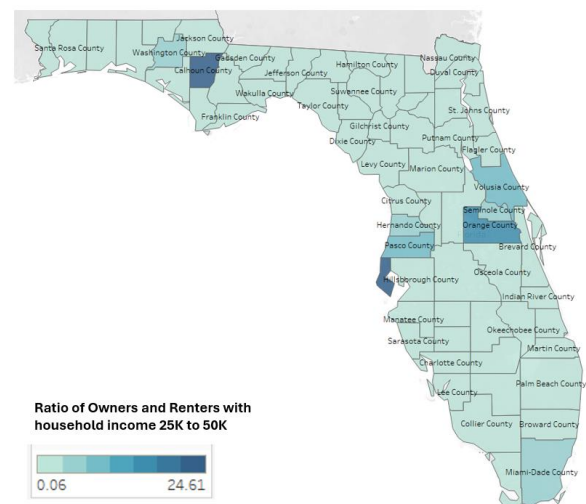


Figure 11 Ratio of Owners and Renters with Household Income 25 to 50K

Middle-Income Households

The closest ACS income grouping that can be considered as middle income, as per Pew Research Center (Kochhar, 2024), is households with income in the range of 50K to 100K. Exploratory analysis on the middle-income bracket can provide insights into the economic status of homeowners. Thus, we created a ratio of renters by owners with household income of 50K to 100K. We used owner household income as the denominator for this calculation as we anticipate more homeowners than renters at the county level for this income group. From Figure 12, it can be noted that Okeechobee and Hamilton

counties have the highest rate of renter over to owners, followed by Dixie and Franklin counties. These top counties have larger rural areas while they have less number of households compared to metro areas, they have a higher relative percentage of renters compared to owners.

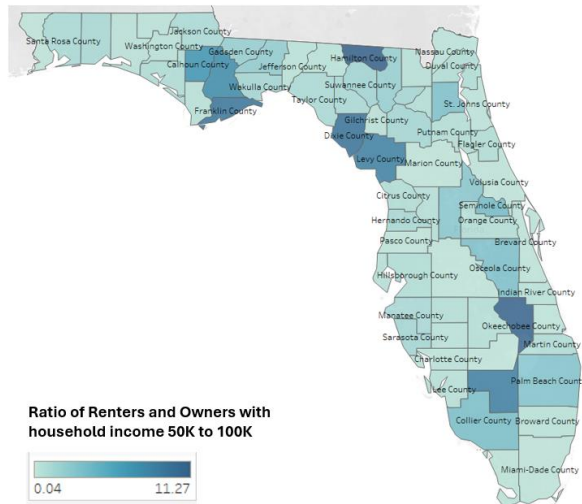


Figure 12 Ratio of Renters and Owners with Household Income 50 to 100K

Renter vs. Owner Vacant Homes and Severe Housing Problems

The higher amount of vacant rental homes and rental properties with severe housing problems can exacerbate the affordable housing issues (Alvarez & Steffen, 2021). In the areas where these problems are higher, rental assistance programs in the form of vouchers could help in reducing worst-case scenarios of affordable housing issues. Rental assistance programs reduce the rent burden to fall within the program limits, and landlords participating in these programs are crucial for addressing affordable housing issues. We gain insights and visualize these factors; we created a scatter plot of substandard rental housing that is vacant against rental houses that have severe plumbing or kitchen facilities problems and cost burden of 50% of income. We create a similar scatter plot for owner based housing. The scatter plots for renters and owner data can be seen in figures 13 and 14.

In Figure 13, we can note that counties that have larger populations and metropolitan cities can be separated from other counties that are grouped in the lower quadrant. Among the larger counties, we can note the problems of vacant rental homes that may have severe problems are quite higher rate in comparison to other counties. A similar

trend can be noted for owners data in Figure 14, with the exception of Lee and Polk counties, which are in the southeast and central regions of Florida.

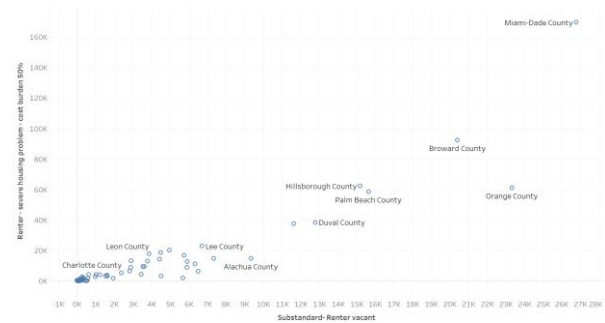


Figure 13 Scatter Plot of Substandard Vacant Rental Homes vs. Rental Homes with Severe Housing Problems

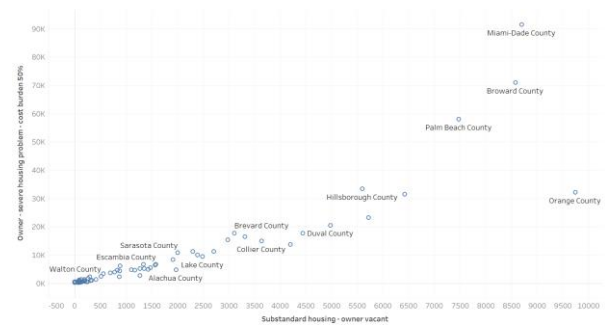


Figure 14 Scatter Plot of Substandard Vacant Owner Homes vs. Owner Living Homes with Severe Housing Problems

5. CONCLUSION

In this paper, we combine and present two related studies wherein the first study's findings encouraged us to proceed with the second study. The first study was a bibliographic review of existing affordable housing literature that investigated the issue using data science methods. We used Google Scholar as a search source. Despite research describing affordable housing increasing in the seven-year search period (2015-2022), articles identified matching our investigation criteria represent a small portion of research published in affordable housing journals. Attempted keyword analysis and comparison of data science-relevant items did not produce any major commonality. The lack of accumulation of coherent literature and scientific findings implies greater investment in data-driven affordable housing research is much warranted. Data science focused research studies could bring up meaningful interventions to effectively and efficiently address the societal challenge of

affordable housing.

It is important to note that this study has limitations in the interpretation of the findings of the research. All analyses on affordable housing literature were created mainly from journal articles. Future studies could consider including papers from reports, books, and conferences. The number of bibliometric records retrieved is limited due to the choice of keywords. Therefore, the keywords used for casting the wide net may not have produced complete coverage of all available relevant literature.

Based on the findings of the first study, we decided to engage in exploratory data analysis at the Florida county level to gain a higher-order level of understanding of the affordable housing issue. During a conversation with one of the nonprofit stakeholders working on the affordable housing issue, they mentioned the disparity between individuals who are homeowners and renters while both groups might face affordable housing issues. The stakeholders informed this as an anecdotal fact, and they have not seen anyone investigate and compare data about those living in affordable housing conditions against those who own a home and are facing affordable housing issues.

We gathered data from the ACS Census and CHAS data sources for the exploratory study. From the analysis, it is clear that the majority of the housing units in Florida are owner-occupied. However, there are significant variations in the ratio of owner-occupied to renter-occupied units in different counties. Some counties have a much higher proportion of renter-occupied units than others. We created several Tableau visualizations of the issue and presented key visuals that shed light on the renters vs. owners scenarios. Visuals indicate very few Florida counties have homeowners for low-income population groups, while rural counties have more renters than homeowners for middle-income population groups. We hope the visualization of the distribution of housing units and the income of renters and owners helps researchers and policymakers create more equitable and sustainable housing policies that benefit everyone in a given community.

This study has several limitations that should be considered when interpreting the findings. First, the data used in this study were collected in 2019 and, therefore, did not reflect the impact of recent events, such as the COVID-19 pandemic, on housing patterns and trends. Second, the data used in this study are self-reported and may be

subject to reporting bias. Third, the sample size for some counties may need to be bigger to draw statistically significant conclusions. Finally, the study is limited to the United States and cannot be generalized to other countries or regions.

There are several avenues for future research in this area. One potential area for future research is to examine the impact of housing policy on the patterns and trends observed in this study. Another potential area for future research is to investigate the role of urban planning and design in shaping housing conditions, ownership, and occupancy.

6. REFERENCES

- Adabre, M. A., & Chan, A. P. C. (2019). Critical success factors (CSFs) for sustainable affordable housing. *Building and Environment*, 156, 203-214. doi:<https://doi.org/10.1016/j.buildenv.2019.04.030>
- Alvarez, T., & Steffen, B. L. (2021). Worst Case Housing Needs 2021 Report to Congress. Retrieved from <https://www.huduser.gov/portal/sites/default/files/pdf/Worst-Case-Housing-Needs-2021.pdf>
- Anderson, L. M., St. Charles, J., Fullilove, M. T., Scrimshaw, S. C., Fielding, J. E., & Normand, J. (2003). Providing affordable family housing and reducing residential segregation by income: A systematic review. *American Journal of Preventive Medicine*, 24(3, Supplement), 47-67. doi:[https://doi.org/10.1016/S0749-3797\(02\)00656-6](https://doi.org/10.1016/S0749-3797(02)00656-6)
- Czischke, D., & van Bortel, G. (2023). An exploration of concepts and policies on 'affordable housing' in England, Italy, Poland and The Netherlands. *Journal of Housing and the Built Environment*, 38(1), 283-303. doi:10.1007/s10901-018-9598-1
- FLHealthCharts. (2024). Florida Median Household Income (Census ACS). Retrieved from <https://www.flhealthcharts.gov/ChartsDashboards/rdPage.aspx?rdReport=NonVitalIndGrp.Dataviewer&cid=0293>
- Galster, G., & Lee, K. O. (2021). Housing affordability: a framing, synthesis of research and policy, and future directions. *International Journal of Urban Sciences*, 25(sup1), 7-58. doi:10.1080/12265934.2020.1713864
- Golubchikov, O., & Badyina, A. (2012). *Sustainable Housing for Sustainable Cities: A Policy Framework for Developing Countries*.

- Retrieved from SSRN:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2194204
- Haffner, M. E. A., & Hulse, K. (2021). A fresh look at contemporary perspectives on urban housing affordability. *International Journal of Urban Sciences*, 25(sup1), 59-79. doi:10.1080/12265934.2019.1687320
- Harzing, A.-W. (2007). Publish or Perish. Retrieved from
<https://harzing.com/resources/publish-or-perish>
- HUD. (2023). Consolidated Planning/CHAS Data. Retrieved from
<https://www.huduser.gov/portal/datasets/cp.html>
- Kochhar, R. (2024). The State of the American Middle Class: Who is in it and key trends from 1970 to 2023. Retrieved from
<https://www.pewresearch.org/race-and-ethnicity/2024/05/31/the-state-of-the-american-middle-class/>
- McCabe, A., Pojani, D., & van Groenou, A. B. (2018). The application of renewable energy to social housing: A systematic review. *Energy Policy*, 114, 549-557. doi:<https://doi.org/10.1016/j.enpol.2017.12.031>
- McCrone, G., & Stephens, M. (2017). *Housing Policy in Britain and Europe*. London: Routledge.

Training a large language model to code qualitative research data: Results from discussions of ethical issues

David Simmonds
davidmsimmonds@gmail.com
Auburn University - Montgomery
Montgomery, AL 36117

Russell P. Haines
hainesrp@appstate.edu
Appalachian State University
Boone, NC 28608

Abstract

Comment coding is an important part of qualitative research, but it is a labor intensive process. Furthermore, researchers need to assess whether or not comments were coded accurately and reliability. Here, we present a process for arranging the original comments and using them to train a Google BERT large language model (LLM) that was able to code comments with 87.9% reliability. This process can be extended by future researchers to potentially code comments made in less-structured research settings, or potentially have the LLM create the comment groupings automatically.

Keywords: transformer model, attention mechanism, text analytics, qualitative research

Recommended Citation: Simmonds, D., Haines, R.P., (2025). Training a large language model to code qualitative research data: Results from discussions of ethical issues. *Journal of Information Systems Applied Research and Analytics* v18, n4, pp 46-55. DOI#: <https://doi.org/10.62273/OTJZ7714>

Training a large language model to code qualitative research data: Results from discussions of ethical issues

David Simmonds and Russel P. Haines

1. INTRODUCTION

In qualitative research, coding data is an important part of the data analysis process (Sarker et al., 2013). Classifying answers to open ended questions is inherently challenging and humans make mistakes and have disagreements about which label to be applied to a given comment (Faraj et al., 2015). Whenever coding of text is involved, raters are expected to operate reliably, meaning they code similar responses in a similar way, every time. The Kappa statistic measures Inter-rater reliability and is used to determine whether there is an acceptable level of match between coders (Haines et al., 2014). An acceptable level of Kappa is considered substantial at .61 according to McHugh (2012).

In this paper, we examine the research question: Can a large language model be trained to code qualitative data in a reliable way? We use the data from Haines et al. (2014) as a training and evaluation set. In that study, they coded comments in discussions about whether actions were ethical or not. Here, the focus is turned to the use of a Large Language model (LLM) for comment coding to determine whether a LLM could provide a comparable reliability to human coders.

In the following, we report both the methods and the results of our LLM training with the idea that other researchers can use the same or similar techniques. Ultimately, our model was able to achieve 87.9% agreement with the human coder, which is quite good considering that the model could not be expected to perform better than the consensus between human beings, which in Haines et al. was 89% with a Kappa of .71.

TRANSFORMERS & the BERT Model

In the seminal paper "Attention is all you need" (Vaswani et al., 2017), Google introduced the Transformer architecture which has revolutionized Natural Language Processing. In 2018, Open AI took the transformer model and split it in half to focus on text generation. This only required the decoder-the second half of the transformer. These foundation models have been pretrained for weeks on thousands of graphical processing units (GPUs) to understand human

language and general knowledge along with some domain knowledge. They have gone on to be embedded in Google Colab for code prediction and Co-pilot in Microsoft Office products. Before transformer models, the AI landscape was dominated by models like the TF-IDF bag of words model for classifying text, or basic neural networks such as RNNs for text generation.

The transformer is made up of an encoder and decoder model. The encoder model makes statistical sense and creates a representation of the patterns and relationships between words and concepts inside the text fed to it. The decoder is then able to generate text from that representation. The big improvement introduced by the transformer model is the ability of the model to remember the relationships between words far away from each other and create more new relationships. Limited only by the size of the input text, transformers can make a map of the strength of the connection between every single word and every other word in the input. This makes them extremely memory hungry, but at the same time, they never forget the relationships. Additionally, transformers are fed word embeddings, which are multi-dimensional vectors of size, 512, 768, 1024 or more. Each position in the vector captures some aspect of the meaning of a word, such that words which mean similar things like dog and puppy have very similar embeddings. Additionally, the difference between words is captured, such that the vectors representing Washington D.C. and USA will have a mathematically similar difference to the vectors representing Berlin and Germany.

For this study, we use Google's BERT transformer model (Devlin et al., 2019). The core of the transformer model is the attention mechanism. There are 3 main types of attention mechanisms: self-attention, multi-head attention and scaled-dot-product attention. Figure 1 shows a diagram which illustrates the attention mechanism and reflects the history and original purpose of transformer models which was language translation. In the diagram, we can see English words with their French equivalent. French and other languages cannot be translated one word at a time, because equivalent words are placed at different relative positions in the sentence. The

attention mechanism allows transformers to create a mapping (weights) which indicate the word in the translated language that the translator should pay attention to when translating words from the original language.

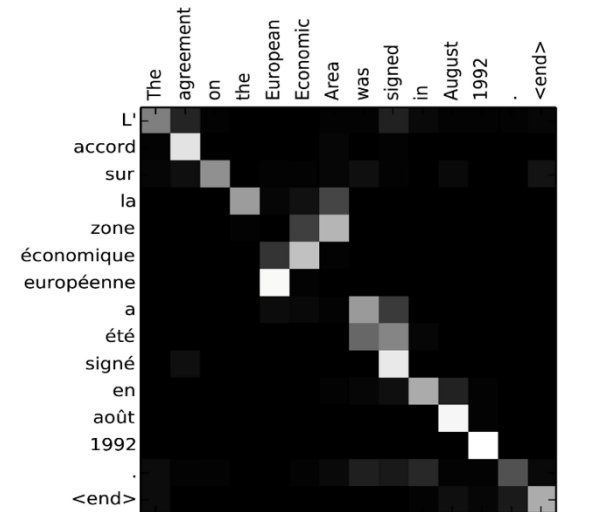


Figure 5: Attention mechanism showing mapping of French and English sentences

An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key (Devlin et al., 2019). The attention mechanism used in the BERT model is called "Scaled Dot-Product Attention" (see Figure 2). The input consists of queries (Q) and keys of dimension (K), and values of dimension (V), and the dot products of the three are computed.

In the case of comment classification, attention similarly tells the classifier which words or phrases to pay attention to in order to predict the correct label for the comment. An example comment in our dataset is "just because it's misleading doesn't mean it's unethical". Before transformers, using techniques like bag-of-words and TF-IDF models, the classifier would see the word "misleading" and interpret it as "unethical". However, the BERT model has been trained on English language and generally knows that any phrase between "just because" and "doesn't mean" implies that the words in between these two phrases do not determine the overall intent of the sentence. And in fact, the speaker usually means the opposite of the words which come after "doesn't mean" – in this case "it's unethical". So, a transformer-based classifier knows to pay less attention to the fact that the action is

misleading because of its model weights; thus, the overall meaning taken from the sentence is the opposite of the next words "it is unethical". These associations (model weights) are then reinforced in the model's fine-tuning on the training dataset since the target label provided in the training dataset indicates that the student was giving support for the action being ethical.

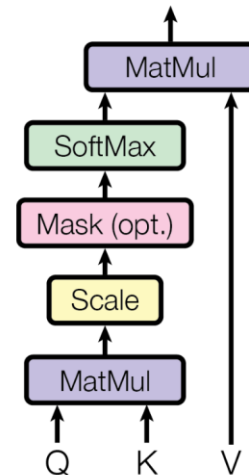


Figure 2: Scaled dot-product attention mechanism reproduced from Vaswani et al. (2017)

2. METHODOLOGY

DATASET

The Haines et al. (2014) dataset was collected during an experimental study of ethical decision making. A total of 219 participants in 42 groups of either five or six members met virtually in chat rooms for three or four minutes to discuss the ethics of five different marketing scenarios. Their final dataset consists of 5,955 "thoughts," which are the comments made by participating students during the online discussions. All of the comments made during the discussions were human coded via a standardized coding sheet according to their contents. Here, we give the codes and their titles, but readers are encouraged to consult the original article if they wish a more extensive explanation of the codes and their definitions. The codes are: supportive remark (SR) stating that the behavior was ethical, a supportive argument (SA) giving reasons why the behavior was ethical, critical remark (CR) stating that the behavior was unethical, critical argument (CR) giving reasons why the behavior was unethical, compromise or accept part of others opinions (CP), neural remark or neural response to other's opinion (NR), agreement on others' opinions (AO), disagreement on other's opinion (DO), query for clarification or explanation (QC), query for

solution (QS), answer to questions (AN), comments on related topic but off track (OT), summary of consensus (SU), uncoded text (UC), humorous comments (HU), off topic comments (OF). Although Haines et al. (2014) reports two coders, only one of which coded all of the comments, the dataset we obtained has three coders, two of which coded all of the comments.

The standard for AI model accuracy is Bayes Error Rate because AI cannot be expected to perform better than a group of human experts who agree. Therefore, comments with mismatched coding were not considered useful. In the original dataset, a few of the comments span over multiple lines of text, meaning that the student broke up a sentence into parts as they participated in the chat. This meant that some of the comments were unlabeled. Comments without a label are also not useful, since there is no ground truth label for the model to be evaluated on. Overall, from our training set, we excluded labels which were empty, uncoded or mismatched between human coders. We dropped those records, leaving 1820 records which had agreement between all 3 coders. 70% of the data (1274 records) were used for the model training, while 30% of the data was used as hold out validation/test sets (273 or 15% each). During hyper-parameter tuning, the validation set was used to determine accuracy of the model after each epoch. This was used for finetuning instead of the model loss since loss does not have a proportional or direct relationship with accuracy. The test set was used to test the final model's accuracy and also generate the predictions.

The code used to denote human coding consisted of the following. The codes themselves had an imbalanced representation of records in the dataset as shown in Table 1. The training set has 312 critical arguments but only 2 that were coded QE, QR, or QP. The coding sheet does not contain either of the last three labels, but looking at the comments themselves, they seem to be miscoded queries that should have been coded either QC or QS. The code used for the label dictionary is as follows:

```
label_dict: {'SR': 0, 'SA': 1, 'CR': 2, 'OT': 3, 'QP': 4, 'CA': 5, 'CP': 6, 'AN': 7, 'QE': 8, 'OF': 9, 'COM': 10, 'QS': 11, 'QR': 12, 'QC': 13, 'NR': 14, 'HU': 15, 'SU': 16, 'AO': 17, 'DO': 18}
```

Rebalancing the dataset improved the accuracy. This came at the cost of stability of the results,

since selecting a different training set and test set tended to change the **validation accuracy** slightly, since prediction of the codes with much less records tended to depend on which of the actual messages were included in the training set. But on average, accuracy increased by about 3% after the rebalancing.

Label	Number of Records
CA	312
EY	259
SA	239
AO	235
CR	209
OF	188
UC	89
NR	80
SR	79
CP	76
QC	39
HU	34
DO	30
QS	30
COM	29
OT	19
SU	15
AN	10
QE	5
QR	2
QP	2

Table 2: Number of records per comment label

The training data was rebalanced by oversampling to become a dataset of 2,000 records with at least 50 records for each label, so the model could get adequate exposure and training to each label. Validation and test sets are not oversampled since they can produce unrealistically high accuracy figures, due to overfitting of the model on the smaller labels which have high consistency due to their repetition.

Columns

The columns used for training included Comment, label, Sequence-# and Scenario-# which was replaced with scenario description. The sequence number was used to reset the sliding window of comments, so that when the sequence # changed to 1, all past comments were erased so that the model would consider only comments related to the particular conversation thread.

Model Training

BERT is a foundation model which can be trained on 2 objectives. The first is next-sentence-prediction. In order to finetune BERT for this objective, it is trained on a dataset which has 2 sentences and a label indicating whether the 2nd sentence follows from the first, or not. The second objective is masked word prediction. BERT can be trained to discover which word is missing (masked) in a sentence in a manner similar to fill in the blank questions given to students on an exam. This makes it suitable for classification of sentences since the comments in this study are responses to the previous comments

Instruction Finetuning

In consumer-facing large language models (e.g., ChatGPT), one can make use of prompt engineering to finetune the results where the model is given a role, objective, output format and reasoning process in order to provide a more suitable prompt completion. Instruction finetuning an LLM is similar to prompt engineering, but it is done on the training side and therefore is a more customized form of finetuning which allows a pretrained model to be customized to a particular task, making the model highly adept at answering questions in the form context expected from users. It learns to think the way a specialist in a profession would think. For example, two students who graduate with a degree in software engineering will have the same knowledge, but if they work as a software developer, trainer, technical writer, or systems analyst their working context will rewire their thoughts to be even more effective at their particular job. Instruction finetuning, combined with Chain-of-Thought reasoning allows models to adapt the knowledge they are given while being trained, showing the way they should use the data to produce the desired output.

Using Chain-of-Thought thought reasoning to means training the model in the way it should think, which can improve its performance significantly. In the case of classification, By including a prompt which tells the model what a the human-being is thinking while labeling the text, allows the model to perform better. (This is also a form of prompt-engineering). In our case, we included instructions indicating that each comment was a response to one of the 5 ethical scenarios, as well as the previous comments. This allowed for great improvements in the model performance. Instruction finetuning, combined with chain-of-thought reasoning allows models to adapt the knowledge they are given while being trained, with the way they should use the data to produce the desired

output. This allows for great improvements in the model performance.

Prompt templates are below:

if Use_Scenario:

```
pt_classify_Sentence1 = ""###  
{Scenario_pretext} \n ### {Scenario} \n\n###
```

```
{Past_Messages_pretext} \n ###
```

```
{Past_Messages}""
```

else:

```
pt_classify_Sentence1 = ""###
```

```
{Past_Messages_pretext} \n ###
```

```
{Past_Messages}""
```

```
pt_classify_Sentence2 = ""###
```

```
{Current_Message_pretext} \n ###
```

```
{Current_Message} \n ### {LLM_Task}""
```

Scenario_pretext = "Below is a scenario given to students. They were asked to comment on whether the scenario was ethical or not"

Past_Messages_pretext = f"Below are {Number_Past_Comments} comments which critiqued the scenario in terms of whether or not the scenario was ethical."

Current_Message_pretext = "Below is the latest comment responding to the previous scenario."

LLM_Task = "Based on the way addressed the scenario and previous comments, create a label for the latest comment."

An example of a hydrated prompt (filled in with data) is shown below in the table below.

SENTENCE-1	Scenario-Preamble	### Below is a scenario given to students. They were asked to comment on whether the scenario was ethical or not.
	Scenario	### On the day welfare checks are received in the ghetto area of a city, a retailer increases prices on all their merchandise for only the store serving that area of the city.
SENTENCE-2	Past-Comments-Preamble	### Below are 2 comments which critiqued the scenario in terms of whether or not the scenario was ethical.
	Past-Comments	### Comment-1: this one is just mean. Comment-2: legal, but unethical.
		### Below is the latest comment responding to the previous scenario.
	Current-Comment	### smart business, but unethical
	Instruction-to-LLM	### Based on the way addressed the scenario and previous comments, create a label for the latest comment.
LABEL	(Label from coder)	1.0

Tokenization

Large Language models cannot process text. They ingest numbers which represent each word (token) in the text fed to it. Hence tokenization is a necessary first step to process text. The dataset was tokenized using the BERT base uncased tokenizer to create model readable tokens. Both model and tokenizer are hosted on Hugging face (www.huggingface.com) and the correct tokenizer is automatically loaded when the checkpoint for the model is used.

Each word or sub-word which can be understood by the model, is part of the tokenizer's vocabulary. Each of these words is represented by a number, from 0 to the vocabulary size (-1). First each sentence is split into words separated by spaces. Sentence1 and Sentence 1 are represented by tokens, numbers which represent each word in the model. When the model predicts a next token, it uses a feedforward output layer with an output head the size of the number of possible predictions. In a text generation LLM, that would be the size of the vocabulary, in the order of 10s of thousands of predictions.

Padding is added to make all the input tokens the same length, since the model can only ingest rectangular batches--having the same number of tokens. Truncation is allowed in order to ensure that sentences that are too long for the 512 token limit are reduced to fit. Sentences are then batched which speeds up training since the GPUs can parallelize the calculations.

Token types are either 0 or 1. 0 indicates that the tokens belong to the first sentence while 1 indicates that the tokens are taken from the second sentence.

Label: is the numeric representation of the code given to the current comment. In this case, since we are only predicting 22 labels, the feedforward layer only has 22 outputs. The tokenizer can only use numbers as labels, so the program creates a dynamic dictionary of codes present in the training dataset. This dictionary automatically changes when different codes are left out of the training set. In order to test the model and create predictions which are human-readable, the dictionary is used to translate the codes back to the original labels.

Input_ids: are the actual tokens representing each word, as described above.

Token-type-ids: allow the model to distinguish between the two sentences fed to it. Every token belonging to the first sentence is assigned a 1. Tokens in the second sentence are assigned a 0.

Attention Mask: indicates which tokens are actual words and which are zeroes added for padding. The model knows to ignore the words used only for padding.

After Tokenization, renaming the label column as 'labels' and removing the columns which will not be used for training, the dataset looks like this:

```
{ 'labels': tensor(5),
```

```

    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
    1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
    0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
    0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
    0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
}

```

The model is then finetuned (trained) on the dataset, over a number of epochs (passes through the dataset). Model training consists of selecting several parameters including Learning Rate, Gamma (rate of reduction of learning rate) and optimizer step size (the number of steps taken down the gradient before learning rate is reduced by multiplying by Gamma).

The model used was a pretrained model from HuggingFace.com. It was downloaded as a model checkpoint: *"google-bert/bert-base-uncased"*. The tokenizer was automatically selected from hugging face using the `from_pretrained` method to ensure compatibility with the model. The size of the model on disk is only 536 MB, but in training, especially depending on the batch size, the memory requirements for the model alone grows 16 to 20 times, easily taking up 10Gb of GPU memory.

Google Colab was used as the coding environment, which has a variable watch window, code completion, code prediction, and code

generation. These features were instrumental in generating code to create a balanced dataset by oversampling the rarer labels (such as QP, QR, QE and AN which only had 2,2,5 & 10 records respectively, compared to CA which had 223 records respectively. The prompt to generate this was:

```
# prompt: split the InstructTuning_df
dataframe into a random sample of 70%
train and 30% validationtest. Then
generate extra records to oversample
the Train set to create a dataset that is
balanced by label. Then split the
validation test into equal parts. Show
the label_column distribution for all 3
resulting datasets.
```

Model Training

Google Colab also gives access to GPUs which are necessary for parallelizing the layers inside the transformer model used by BERT. An L4 GPU on Google Colab with 22Gb of GPU memory was able to accommodate 64 records per batch of training data, training the model in 25 seconds per epoch. With 20 epochs taking between 9 and 10 minutes.

The concept of LLM "loss" is based on statistics, being the difference between a set of predictions made by the model and the actual values in the dataset. It is founded in in mathematical optimization and statistics. In statistical regression, the software tries several lines of best fit to match the datapoints. It then calculates loss using ordinary least squares (OLS) and accepts the solution with the lowest OLS. Training an LLM is very similar, since the training objective is to minimize the model loss. In LLM, the gradient descent is used to achieve the minimal loss-or best model. At the start of training, model weights (similar to regression coefficients) are randomly initialized. In each of the several "epochs", The model uses weights to generate predictions which are compared to target labels. The loss is then calculated and backward propagation occurs, in which a fraction of the loss is subtracted from the model weights. Trial and error is required to find the optimal hyperparameters (learning rate, gamma, step size). Several runs need to be done in order to determine the best combination of batch_size, epochs, learning rate, gamma and optimizer step size.

The model was eventually trained for 7 epochs with a learning rate of 5.1 e-5, optimizer step size of 1 and gamma of .81. We arrived at these parameters after adjusting the learning rate from 7, 6, 5, 4, 5.5 with several more tweaks and

finally 5.1 (e-5). Gamma started at .99 and went to 0.91, 0.8, 0.7, 0.85 with several more adjustments and finally .81. The training accuracy over training epochs is shown in Figure 3 below. The ShedulerLR optimizer was used, which allows for a decay in the training rate as it approaches the lowest loss. The learning rate needs to be sufficiently large in the beginning to achieve high accuracy. The loss decay is necessary because as the model approaches its lowest possible loss, it needs to creep to the minimum loss without overshooting. Hence an exponentially decreasing learning rate is required as provided by StepLR. The learning rate is multiplied by $\text{Gamma}^{\text{TrainSteps}}$ and since Gamma is < 1 , the learning rate approaches zero. For example, with a Gamma of .82, on the 20th training step the learning rate would become $0.82^{20} = 2\%$ of the initial learning rate.

```
bert-base-uncased      Number of sentences: 3
Learn_rate: 5.15e-05,  Gamma: 0.81,  Optim_steps: 1
*****
epoch: 0,  27.3s,  loss = 1.5880,  Accuracy = 52.0%
epoch: 1,  26.8s,  loss = 0.7721,  Accuracy = 83.2%
epoch: 2,  26.4s,  loss = 0.4383,  Accuracy = 83.5%
epoch: 3,  26.1s,  loss = 0.1140,  Accuracy = 87.5%
epoch: 4,  26.5s,  loss = 0.2578,  Accuracy = 87.2%
epoch: 5,  26.5s,  loss = 0.1119,  Accuracy = 87.2%
epoch: 6,  26.4s,  loss = 0.0179,  Accuracy = 87.9%
```

Figure 3: Training accuracy and gamma over training epochs

3. RESULTS

After training, a hold out dataset was used to test the overall accuracy of the model and generate predictions which could be compared by humans against the ground truth labels. A predicted label as well as a prediction confidence was produced. This allowed us to differentiate between comments that were labeled incorrectly with low confidence vs those with high confidence. We achieved a finetuned model with a training loss of .051 and validation accuracy of 87.9%. The model was used to predict the labels on our test set. González-Carvajal, & Garrido-Merchán (2020) categorized Portuguese News items and achieved accuracy of 91%. The first dataset had 9 categories, equivalent to ours, but most articles had between 300 and 700 words, significantly longer than our messages. They also achieved a score of 83% on the categorization of tweets into real or non-real disasters. In the 2nd case, they used 10875 tweets multiples of our dataset and with only 2 categories compared to our 9.

4. LIMITATIONS AND CHALLENGES

The study would have benefited a lot from more data. Studies of this type typically have many

thousands of rows of data, for example González-Carvajal et al (2020) which used 167,000 and 10,800 records. By increasing the data available, the classifier would be able to learn about the patterns. We had 5000 records but excluded the ones that the human beings did not agree on. This could have been an avenue of improvement. The length of the messages is also a big problem. Many messages were as short as "OK" which gives practically no indication what they were responding to. Possibly even more challenging is the fact that the messages were captured in a continuous stream, not as responses to any particular previous message. It is very possible for a message to be a response to one 20 messages ago, depending on typing speeds.

5. DISCUSSION AND CONCLUSIONS

The ability of any model to classify a set of comments will be heavily influenced by the initial training data. It means researchers should try to collect training data about many scenarios, from many different sources. A meta study would then be able to remove local and situation biases in attitudes to ethics from affecting the study.

LLMs are generally trained with the objective of next word prediction by minimizing the loss between predicted words and actual words given to it in a training dataset. BERT is special in that it is trained to also do next sentence prediction: whether a second sentence follows from a first. To allow BERT to achieve this, a special classification layer is added to the other layers to allow for classification (Devlin et al., 2019). The model can then be finetuned on pairs of sentences as in our case. Though the foundation model was trained with a classification head that produces a 0 or 1: whether or not sentence-2 follows sentence-1, it can be finetuned to produce many classes, given the right number of training data. In our case, this was 9 classes.

Whereas a regular LLM uses an input sentence to predict the next word, our code and configuration meant BERT was essentially using 2 sentences to produce a next word – which in our case is the comment label. Another analogy might be a conversation between a person and an LLM: the person says "Hi", the regular LLM would be limited to saying "Hello". Our BERT model takes this a step further to act as an observer between the two people in the conversation. Supposing it is finetuned on a conversation between two polite speakers. It is then asked to assess 2 sentences -- Sentence-1: "Hi, how are you?", Sentence-2: Hello I am well, how about you?" BERT would predict a "1" (yes 2nd sentence follows from the

1st). If it were fed a conversation in which one impolite person is involved. Sentence-1: "Hi, how are you?", Sentence-2: "We need to go to lunch right now because I am hungry?" it should predict a 0 (no 2nd sentence does not follow from the 1st because some pleasantries should precede the 2nd response).

In conclusion, we are confident that the answer to our research question is that an LLM can indeed be used for comment coding. Our accuracy of 87.9% would be somewhat convincing if we were trying to predict 2 labels, but is even more convincing because we are in fact predicting 9 labels. We were able to take this step by building on Adhikari et al. (2019) and finetuning BERT to predict more than 2 classes, by asking the BERT model: *How* does the second sentence follow the first? Is it agreement, disagreement, neutral, critical or supportive? This code and configuration allowed us to label 9 different classes of comments with a substantial level of reliability, before any finetuning measures are even applied. Future researchers can build on these results by examining other comment coding situations. For example, it might be possible to code more categories, longer and less-focused comments, or to even have the transformer model define the classifications itself.

6. FUTURE RESEARCH

There are some things we can attempt in order to improve the accuracy of the classifier. There is the possibility of looking at the actual codes used to classify. It became apparent after a quick review, that there were codes which were confused with each other, and the realization that a human would have probably confused them too. We also plan to explore ways in which the classifier could actually challenge the human ratings. That is to say, using generative AI to explain why it chose a particular code. So, it would not just be right or wrong, just more or less reasonable.

We plan to address the empty and mislabeled comments in a future study. We will look where we address the creation of an optimal set of labels which avoid the overlap with other labels and the gaps between labels. But for this study, we decided to limit the discussion to the use of BERT to assist human coders.

7. REFERENCES

Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). *DocBERT: BERT for Document Classification*

- (arXiv:1904.08398). arXiv.
<http://arxiv.org/abs/1904.08398>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv.
<http://arxiv.org/abs/1810.04805>
- Faraj, S., Kudaravalli, S., HEC Paris, & Wasko, M. (2015). Leading Collaboration in Online Communities. *MIS Quarterly*, 39(2), 393–412.
<https://doi.org/10.25300/MISQ/2015/39.2.06>
- Haines, R., Hough, J., Cao, L., & Haines, D. (2014). Anonymity in Computer-Mediated Communication: More Contrarian Ideas with Less Influence. *Group Decision and Negotiation*, 23(4), 765.
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282.
- Sarker, S., Xiao, X., & Beaulieu, T. (2013). Guest Editorial: Qualitative Studies in Information Systems: A Critical Review and Some Guiding Principles. *MIS Quarterly*, 37(4), iii–xviii.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.
<https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

AI-Related Advertising on Facebook: Addressing Bias, Targeting Challenges and Regional Factors

Sera Singha Roy
sera.sroy@gmail.com

Tanya Linden
lindent@unimelb.edu.au

University of Melbourne
Parkville, VIC 3010 Australia

Abstract

It has been established by research and marketing firms that advertising on social media is due to its reach, cost effectiveness and range of media formats appealing to a range of target audiences. However, a campaign needs to be properly designed to reach the intended audiences and therefore be value for money. Therefore, the focus of this research is to evaluate the effectiveness of AI-related advertising campaigns, particularly in reaching their intended audiences, such as individuals with a declared interest in AI. This study addresses the under-explored area of how effectively these campaigns target and engage their intended demographic groups, revealing gaps and opportunities in current advertising strategies. A dataset of 5,528 AI-related advertisements was analyzed to evaluate the parameters of targeting precision, demographic distribution, and geographical variations. The analysis aimed to identify discrepancies in ad reach and effectiveness across different demographic segments and various geographical regions. The findings show a significant discrepancy between overall ad volume and targeted reach, with only 12.6% of ads directly targeting individuals interested in AI. Our findings reveal a distinct demographic skew in AI ad exposure, favoring males and the 25-34 age group. This gender disparity is particularly pronounced when targeting AI enthusiasts, raising concerns about potential bias in ad delivery. Significant geographical variations in ad performance were also identified, underscoring the importance of regional context in advertising strategies. The implications of these findings for advertisers are clear: more precise targeting is needed to engage AI-interested audiences effectively, algorithmic biases must be addressed to ensure equitable ad delivery, and contextual factors should be considered to optimize campaign strategies.

Keywords: Digital Marketing, Social media advertising, Social media analytics, Data Analytics, Facebook Analytics, Artificial Intelligence

Recommended Citation: Singha Roy, S., Linden, T., (2025). AI-Related Advertising on Facebook: Addressing Bias, Targeting Challenges and Regional Factors. *Journal of Information Systems Applied Research and Analytics*. v18, n4, pp 56-66. DOI# <https://doi.org/10.62273/IJSS6362>

AI-Related Advertising on Facebook: Addressing Bias, Targeting Challenges and Regional Factors

Sera Singha Roy and Tanya Linden

1. INTRODUCTION

In today's contemporary digital landscape, advertising on social media plays a crucial role in modern marketing serving as a powerful tool for targeted advertising, customer engagement, and brand loyalty (Lulandala, 2022). Social media platforms like WhatsApp, Facebook, and YouTube offer advertisers the ability to reach potential customers 24/7, leveraging the platforms' extensive reach and user engagement. This in turn opens vast opportunities for brands to connect with consumers, build brand recognition, and ultimately drive sales (Mishra, 2017; Vitalis et al., 2023). The ability to precisely target demographics, track user engagement, and measure campaign effectiveness has solidified social media advertising as an essential element in the modern marketing mix (Sharma & Ashfaq, 2023). However, the effectiveness of social media advertising hinges on a deep understanding of demographic parameters and their profound influence on consumer behavior.

When considering platforms that excel in targeted advertising, Facebook emerges among the best choices due to its massive user base coupled with sophisticated targeting capabilities and diverse ad formats, providing businesses with optimal opportunities to reach potential customers (Kawaf & Istanbuluoglu, 2019). Advertisers can utilize Facebook's detailed and diverse targeting options to connect with specific demographics, interests, and behaviors, ensuring that their ads are seen by the most relevant audience (Risius & Ayding, 2018). Additionally, Facebook's advanced analytics tools provide marketers with capabilities to monitor campaign performance, and make data-driven decisions to optimize their advertising strategies (Adeleye et al., 2024).

Facebook's advertising reach extends to a wide array of fields, including healthcare, sustainability, entertainment, education, and even cutting-edge Artificial Intelligence (AI) tools. As AI is rapidly transforming various industries such as healthcare, finance, education, retail, and banking to name a few (Mungoli, 2023), we are observing a surge in advertising campaigns promoting AI products, services, and content (Marinchak et al., 2018). Throughout this study,

the term "*AI-related advertising*" will refer to any advertisement that features AI products, services, or content. Examples of AI-related advertisements could range from an online competition to complete an AI coding challenge to a healthcare service utilizing AI chatbot services to help a patient recognize the seriousness of initial symptoms. However, the effectiveness of these campaigns in reaching their intended audiences remains underexplored. There is a need to better understand the optimal use of targeting strategies, the potential for algorithmic bias, and the influence of geographical and cultural factors on these campaigns for businesses to get the best outcomes and value for money (Whitelegge, 2016). This study investigates the targeting precision and demographic reach of advertising campaigns featuring AI-related content, as well as examining factors influencing campaign effectiveness across different demographics and geographical regions. The following research questions were posed.

RQ-1: To what extent does interest-based targeting influence the reach and effectiveness of AI-related advertising compared to other targeting strategies?

RQ-2: How do advertising strategies and broader societal factors intersect to shape the representation of different demographic groups within AI-related advertising?

RQ-3: How do geographical and contextual factors influence the performance of AI-related advertising campaigns across different regions?

2. BACKGROUND

Facebook advertising effectiveness is a multifaceted landscape, influenced by user demographics, ad content, and engagement metrics (De Silva, 2023). Research consistently demonstrates its impact on purchasing behaviors, particularly among younger demographics. For instance, research conducted in the Klang Valley area of Malaysia found a strong correlation between Facebook's influence and the effectiveness of digital advertising among youth, with socio-demographic factors such as age and

educational level playing a role in this relationship (Lin et al., 2024). Studies also highlight the positive correlation between Facebook's influence and digital advertising effectiveness, with sociodemographic factors playing a significant role. This effectiveness extends to national campaigns, where engagement metrics like likes, comments, and shares are key performance indicators, informing strategies for brand image enhancement and customer acquisition (Kolesnyk & Kostynets, 2023).

However, while research on Facebook advertising effectiveness is extensive, studies specifically examining campaigns involving AI-related products and services are limited. This gap is critical to address, given the unique challenges and opportunities presented by this rapidly evolving technological landscape.

Demographic parameters, such as age, gender, income, education, and occupation provide crucial insights into the specific needs, preferences, and purchasing behaviors of different customer segments (Dobscha & Ostberg, 2021). Gender is important as a social indicator of what is considered an expected role and behavior within the discussed context (Birknerová et al., 2018). It is also associated with prejudices and stereotypes, however these change with time. Research indicates that men and women have different perception of advertising materials, as well as different expectations in this area (Birknerová et al., 2018). It has been evident that income levels can influence responsiveness to marketing, with higher-income individuals potentially favoring luxury and exclusivity, while lower-income groups may prioritize affordability (Patino et al., 2014). In the healthy food market, for instance, socially responsible marketing practices are seen as long-term investments that enhance company value, with key demographic variables like gender and household income significantly affecting customer perceptions (Vržina & Pepur, 2021). Age groups must be considered during the design of marketing campaigns because there are significant differences in cognitive functions of young consumers compared to mature age consumers. As a result, different age groups behave differently in relation to persuasion delivered by advertising. Younger consumers are more skeptical of advertising whereas older audiences are more susceptible to promotional tactics (Phillips & Stanton, 2004).

Understanding of these parameters is paramount for crafting effective marketing strategies across various domains, including targeted marketing

campaigns, product development, pricing strategies, distribution channels, and customer relationship management (Dobscha & Ostberg, 2021). Geographic parameters further influence social media marketing by adding layers of cultural nuances, local trends, and communication styles (Patino et al., 2014). Local trends and consumer preferences, often varying widely across regions, necessitate adapting social media strategies to reflect these nuances. The integration of geographic data with social media networks allows marketers to visualize and identify potential local connections between firms and customers, enhancing targeted marketing efforts (Fogarty, 2021). Finally, language and communication styles prevalent in different geographic regions should be carefully considered, ensuring messages are not only translated accurately but also culturally adapted for resonance with the local audience (Patino et al., 2014).

Previous research has highlighted the potential for algorithmic bias, particularly concerning underrepresented or marginalized groups. For instance, studies have found that online ad delivery can perpetuate gender stereotypes, with women more likely to see ads for caregiving roles and men for leadership positions (Rodgers & Nguyen, 2022). Similarly, research has uncovered racial biases in the delivery of housing and credit-related ads, with ads for arrest records more likely to be shown to Black users (Gao et al., 2023). This is particularly crucial for emerging technologies like AI, which carry their own set of societal implications and its impact on various sectors (Chander, 2024).

Furthermore, geographical and cultural context significantly influence advertising effectiveness. Scholars emphasize understanding local market dynamics, consumer behavior, and cultural nuances when developing global advertising strategies (Yu, 2021). In Saudi Arabia, for instance, global brands that incorporate local cultural cues related to Islamic values and family dynamics in their Facebook content see higher consumer engagement, illustrating the importance of cultural value congruence in social media marketing (Abuljadail & Badghish, 2021). This is particularly relevant for AI-related advertising, as the adoption and perceptions of AI technology can vary widely across different regions and socioeconomic backgrounds (Kelley et al., 2021).

Therefore, drawing upon existing research on targeted advertising, particularly on platforms like Facebook, is essential for framing the

investigation of AI-related advertising campaigns. This context highlights the need to examine potential biases, consider cultural nuances, and understand the unique challenges and opportunities presented by this rapidly evolving technological landscape through our study.

3. METHODOLOGY

The study employed quantitative research approach, which encompasses distinct stages devised to extract, process, analyze, and interpret the gathered data to discern nuanced insights into AI-related advertising on Facebook.

Data Collection

The data collection process involved accessing and extracting pertinent details from two primary datasets obtained from Facebook's developer platform, specifically Meta's Research Platform: the Ad Targeting dataset and the Ad Library dataset. These datasets provided comprehensive insights into the targeting criteria, ad creatives, and performance metrics of advertisements related to AI.

The Ad Targeting dataset (`ad_targeting_dataset_sieep_aug_2020`) (Meta Platforms, 2024) provides detailed information about the targeting criteria used for each advertisement on Facebook, including demographic parameters such as age, gender, location, and interests. These options are provided to the advertisers when they are creating their advertisement and specifying targeting criteria. The 'interest' field for each user is compiled by Facebook from pages liked by the user or from keywords associated with pages or apps the user uses. Data collection from this dataset began with the formulation of SQL queries designed to filter ads targeting interests related to AI. These queries were tailored to include specific keywords and phrases relevant to AI such as 'A.I.', 'Artificial Intelligence', 'Deep Learning' and so on. The Ad Library dataset (`ad_archive_api`) contains information about ad creatives and performance metrics, including

- ad creative body – texts that are displayed in an ad,
- spend – amount of money spent on running an ad,
- reached countries – countries that ran an ad
- regional distribution– regional distribution of people reached by an ad,
- demographic distribution – demographic distribution (age range and gender) reached by an ad,
- impressions – number of times an ad has been on the screen of users

Integration of the Ad Targeting dataset with the Ad Library dataset was facilitated using unique identifiers such as archive IDs, allowing for the correlation of targeting criteria with ad performance metrics. Data extraction from this dataset involved querying ads that were identified in the Ad Targeting dataset using prominent keywords related to AI such as 'A.I.', 'Artificial Intelligence', 'Deep Learning' and other synonyms. For instance, the following query provides us with the unique archive id and demographic distribution for each ad which are targeted to users interested in 'Artificial intelligence'. The column 'ds' represents a date stamp (e.g., 2021-11-03).

```
SELECT a.archive_id,  
       b.demographic_distribution  
FROM {database}.{table} a,  
     {database}.{api_table} b  
WHERE a.ds = b.ds AND a.archive_id =  
       b.fb_id  
AND CARDINALITY( FILTER(  
  CAST(JSON_EXTRACT(include, '$') AS  
  ARRAY(MAP(VARCHAR, VARCHAR))), (x) ->  
  ELEMENT_AT(x, 'Artificial  
    intelligence') = 'Interests')) > 0
```

Leveraging the expansive scope of the Ad Targeting dataset, which encompasses detailed targeting criteria utilized by advertisers, and the Ad Library dataset, containing comprehensive performance metrics, a robust dataset gathered from April 21, 2021 was compiled for subsequent analysis.

Data Preprocessing and Cleaning

Upon acquisition of raw data from the Ad Targeting and Ad Library datasets, data cleaning steps were undertaken to ensure the integrity and consistency of the dataset. These endeavors aimed at rectifying inconsistencies and addressing missing values. Using extensive SQL queries, the 'NULL' and missing values from the columns were removed to minimize data anomalies and enhance the overall quality of the dataset and therefore, the results.

Data Analysis and Interpretation

A diverse array of analytical techniques was employed to extract meaningful insights from the processed dataset. Key performance metrics, including impressions, interest in AI, age, gender, and location, were subjected to comprehensive analysis to discern patterns, trends, and correlations. Leveraging both queries in SQL and visualization of the insights through graphs and plots, this research aimed to identify key patterns and relationships within AI-related advertising.

This study examines the influence of targeting strategies on the performance of AI-related advertising across various dimensions. We analyzed a final dataset of 5,528 AI-related advertisements to understand how interest-based targeting, demographic representation, and geographical context shape advertising outcomes. Our findings provide insights into the effectiveness of different targeting approaches and highlight potential biases embedded within current advertising practices.

RQ-1: To what extent does interest-based targeting influence the reach and effectiveness of AI-related advertising compared to other targeting strategies?

Our findings demonstrate that interest-based targeting significantly impacts the reach and effectiveness of AI-related advertising. A mere 12.6% of the analyzed ads (700 out of 5,528) were targeted based on consumers with declared interest in AI. However, these targeted ads achieved an average of 45,275 impressions, substantially surpassing the 28,764 impressions for ads targeting individuals without a declared interest in AI and the 30,855 impressions for ads targeted to audience irrespective of their interests as shown in Figure 1.

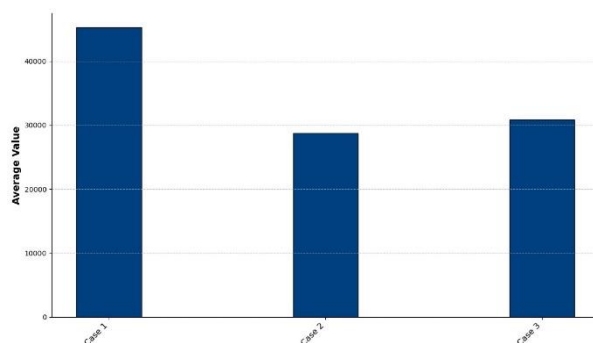


Figure 6 Comparison of impressions made by AI-related advertisements for targeted audiences. Legend: Case 1 - Ads targeted to audiences interested in AI, Case 2 - Ads targeted to audiences not interested in AI and Case 3 - Ads targeted to audience irrespective of their interests

Gender disparities were evident, with males representing a disproportionately larger share of the targeted audience (62.98% overall, 65.80% for targeted ads) compared to females (36.21% overall, 34.17% for targeted ads) as depicted in Figure 2 and Figure 3. The age group 25-34 exhibited highest exposure to AI-related ads with

or without considering the audience interest.

The 18-24 age group shows a significant increase in exposure when advertisements are catered with respect to audience interest. This indicates a greater likelihood of engagement with AI-related content of ads when perceived relevance is high. Conversely, the 45-54 age group experiences a substantial increase in exposure when ads are not targeted as per audience interest. This observation presents an intriguing paradox: while this demographic may not actively express interest in AI, their passive exposure through untargeted campaigns appears to be substantial (Barbul & Bojescu, 2023). This could indicate an underlying receptivity to AI-related advertisements that is not captured through explicit interest-based targeting (Beauvisage et al., 2023).

Therefore, the results for RQ1 suggests that aligning advertising content with audience interests is crucial for enhancing campaign effectiveness (De Maio et al., 2020). Interest-based targeting holds significant potential for engaging a more receptive audience and maximizing the impact of AI-related advertisements (Sharma & Ashfaq, 2023).

RQ-2: How do advertising strategies and broader societal factors intersect to shape the representation of different demographic groups within AI-related advertising?

Analysis of demographic data reveals potential biases in the delivery of AI-related advertising. From Figure 4 we can observe that across all ads, the 25-34 age group exhibited the highest impression (28.7%), followed by the 18-24 (22.2%) and 35-44 (18%) age groups. Although 25-34 age group depicts highest impression on AI-related ads, a lot of potential in achieving a higher impression could be seen in the 18-24 and 35-44 age groups. Because of the gender disparity in ads targeting audience as discussed in RQ1, gender biases in impressions are observed due to the lower exposure of AI-related ads towards females. These findings raise concerns about potential biases embedded within ad delivery algorithms and strategies, thereby highlighting the broader representation issues within the AI industry (Kaplan et al., 2022; Lambrecht & Tucker, 2019). Both targeting methodologies and societal factors likely contribute to these disparities (Lambrecht & Tucker, 2019).

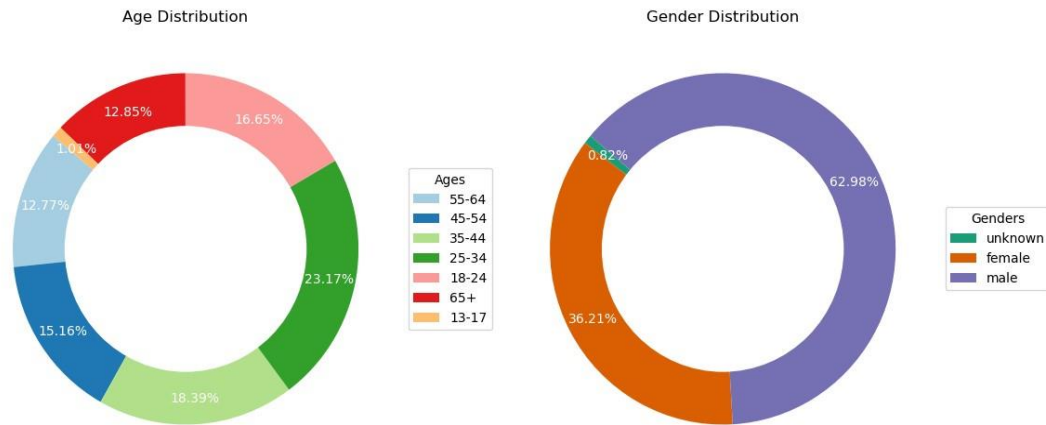


Figure 2 Demographic distribution of AI-related advertisement

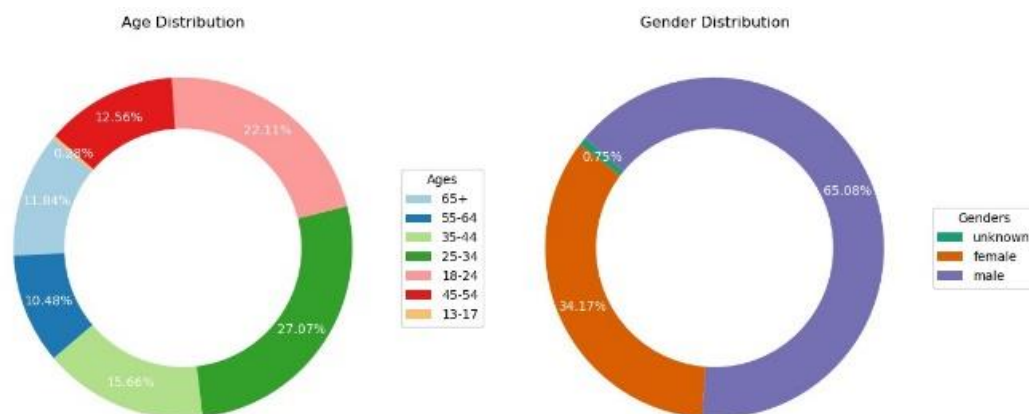


Figure 3 Demographic distribution of AI-related advertisement targeted towards interested audiences

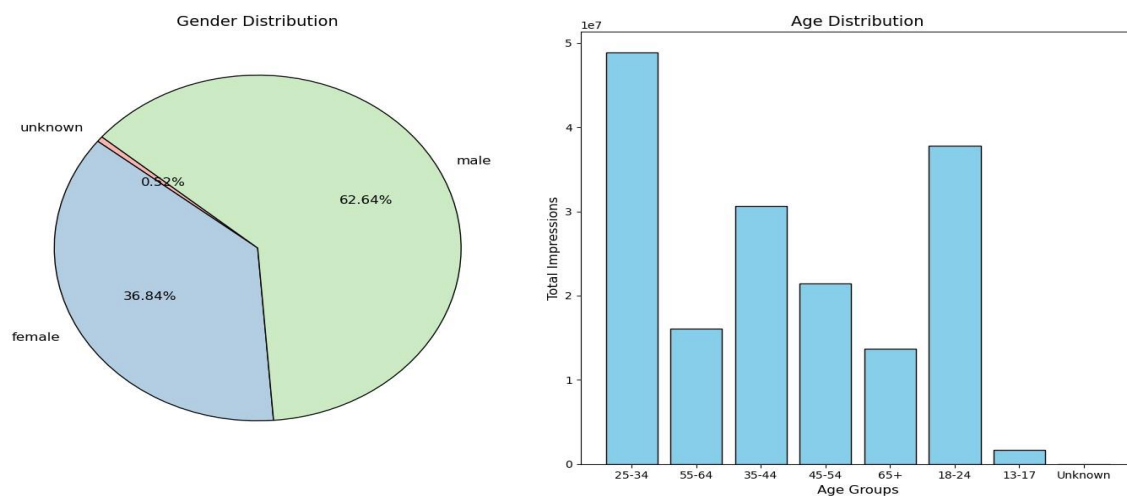


Figure 4 Depiction of impressions of AI-related ads over gender and age distributions

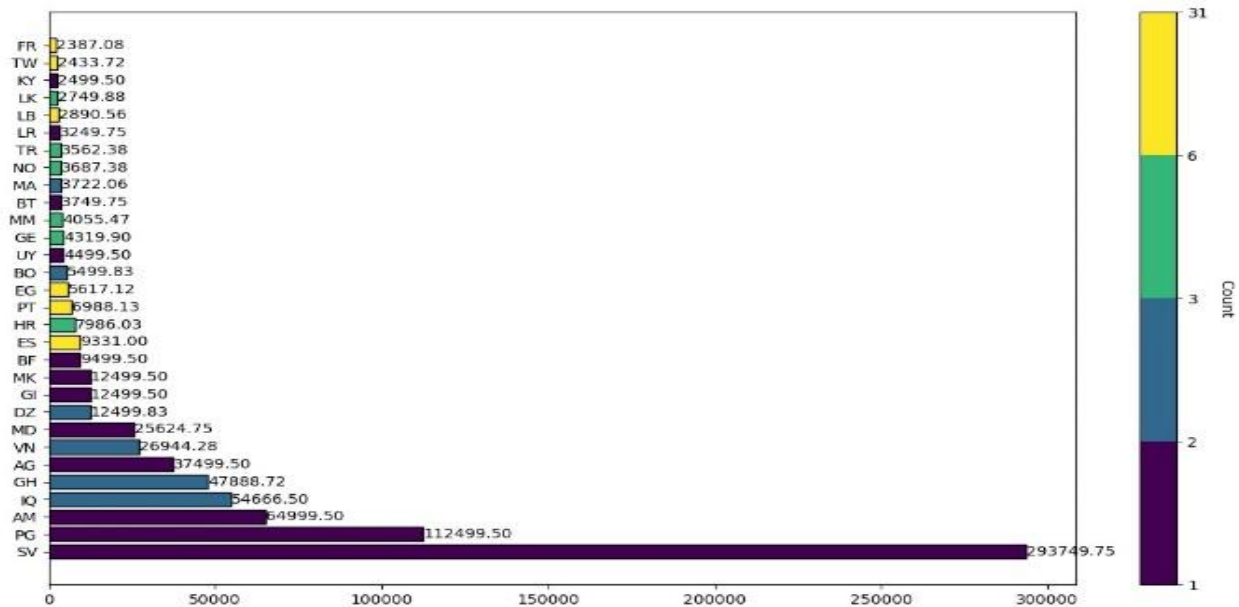


Figure 5 Top 30 countries (ISO Alpha-2 code) with highest impressions per ad

RQ-3: How do geographical and contextual factors influence the performance of AI-related advertising campaigns across different regions?

Geographical analysis reveals significant variations in ad impressions and reach across different countries. For instance, from Figure 5 we can observe that El Salvador and Papua New Guinea exhibited high impressions per ad but a lower overall ad count, suggesting a strategy focused on maximizing reach within a smaller, potentially niche audience. Conversely, France demonstrated lower impressions per ad but a higher ad count, indicating a focus on repeated exposure within a larger, more competitive market. These findings emphasize the importance of contextualizing ad performance data within the specific characteristics of each geographical region (Kannan, 2017). A nuanced, context-aware approach is essential for developing and implementing effective ad campaigns (Anderson et al., 2019). Advertisers must tailor their strategies based on regional market dynamics, understanding local contexts and audience behaviours to optimize ad performance (De Maio et al., 2020).

5. DISCUSSION

Interest based advertising implications

In answering Research Question 1, our findings underscore the significant impact of interest-based targeting on the reach and effectiveness of AI-related advertising. Although only 12.6% (700

out of 5,528) of the ads directly targeted individuals with a declared interest in AI, these ads achieved substantially higher impressions (an average of 45,275) compared to untargeted ads (30,855 impressions) and ads targeting individuals without a declared interest in AI (28,764 impressions). This demonstrates that targeted audience tend to view the advertisements featuring AI more frequently than untargeted audiences. This finding supports the importance of precise targeting in reaching audiences genuinely interested in AI-related products and services in line with the previous research (De Silva, 2023; Kawaf & Istanbuluoglu, 2019), emphasizing the multifaceted effectiveness of Facebook advertising influenced by user demographics and engagement metrics. Further research could benefit from other forms of engagement metrics such as sharing or click-rates.

However, the observed gender disparities in advertising reach raise concerns about potential algorithmic and strategic bias. Males represented a disproportionately larger share of the audience both overall (62.98%) and for targeted ads (65.80%) compared to females (36.21% overall and 34.17% for targeted ads). This finding is also consistent with past research which indicated that targeted advertising can perpetuate societal inequalities (Rodgers & Nguyen, 2022). The overrepresentation of males in AI-related advertising exposure, even when targeting those interested in AI, suggests that algorithmic

systems may amplify existing gender imbalances within the AI industry, as observed by (Lambrech & Tucker, 2019). This finding also aligns with the study by (Kolesnyk & Kostynets, 2023), who highlighted the role of socio-demographic factors in digital advertising effectiveness.

It is important to point out that interest in AI in Ads Targeting Dataset relies on user engagement in the form of keywords or pages liked by users. Although the advertisers have an option to select their target audience based on their age, gender, interests, and behaviours, the "interest in AI" can vary across any age group or gender. For instance, an advertisement on healthcare enhancements through incorporating AI might not be appealing to a student who is healthy and learning a second language so their interests would be concentrated on other AI capabilities. Advertisements promoting new AI enhanced language courses would spark more interest in that user. Although demographic data and interest categories offer a reliable starting point, more robust approach involves considering contextual relevance, engagement with specialized communities and forums, content consumption patterns related to AI, and targeting based on self-reported skills. By incorporating these strategies, advertisers can reach more receptive audiences while respecting user privacy and fostering a more relevant and engaging online experience.

Demographic Representation and Algorithmic Bias

The demographic analysis, as per Research Question 2, further substantiates concerns about biases in AI-related advertising. The overrepresentation of the 25-34 age group (28.7% of impressions) and the underrepresentation of females across all ad categories suggest that both algorithmic and broader societal factors contribute to these disparities. This mirrors the findings of Lin et al. (2024), who found that socio-demographic factors such as age and educational level significantly influence digital advertising effectiveness among youth.

Research indicates that AI-advertising algorithms, potentially trained on historical data reflecting existing biases, may inadvertently perpetuate these disparities (Gao et al., 2023). This identified problem highlights the need for increased transparency and accountability in the development and deployment of such algorithms, particularly given the ethical implications of promoting emerging technologies like AI. This concern is echoed by (Kaplan et al., 2022), who

advocate for ethical considerations in algorithmic decision-making processes.

As observed, the impact of algorithmic bias could lead to underrepresentation of certain demographic groups, therefore it is crucial for social media platforms to develop strategies to mitigate these biases. Suggesting potential mitigation methods requires significant transparency from social media platforms as well as enhancement of the current algorithms and development of new ones (which is beyond the scope of this study).

Geographical Context and Advertising Performance

The geographical analysis, as performed to answer Research Question 3, reveals the significant influence of context on AI-related advertising performance. Variations in ad impressions and reach across different countries highlight the need for tailored advertising strategies that consider local market dynamics, cultural nuances, and varying levels of AI adoption. For instance, El Salvador and Papua New Guinea exhibited high impressions per ad but a lower overall ad count, suggesting a strategy that focused on maximizing reach within a smaller, potentially niche audience. Conversely, France demonstrated lower impressions per ad but a higher ad count, indicating a focus on repeated exposure within a larger, more competitive market. These contrasting strategies demonstrate the importance of understanding regional differences in audience size, engagement patterns, and advertising saturation (De Maio et al., 2020). This emphasis on regional adaptation resonates with the findings of (Patino et al., 2014), who stressed the importance of considering cultural nuances, local trends, and communication styles in social media marketing. This aligns with the study by (Yu, 2021), who emphasized the need for culturally sensitive approaches in global advertising campaigns.

Practical recommendations for Advertisers

- Leveraging the interest-based targeting option - Facebook Ad Targeting Dataset provides options to the advertisers to select their audience from a range of demographic groups and interests. As pointed earlier in the paper, majority of advertisers do not specify their target audience based on their interests. Therefore, advertisers should align their ads with the audience with declared interest in AI.
- Opting inclusive strategies - Along with the most deserving group with declared interest, the advertisers could potentially target the consistently underrepresented groups

exhibiting relevant interests, thereby providing a pathway for the ads to reach them. The advertisers can also consider various ethnicities, gender and age groups in ad images and ad creative body. This will address the biases as well as a larger audience would be able to resonate with the ad.

- Localized campaigns – The advertisers should focus on adapting their advertising strategies with respect to the local trends rather than generalizing strategies for all regions.

Implications for Platform Developers

- Engagement with advertisers – Understanding from the viewpoint of advertisers will help the platform to understand their needs and therefore, develop bias mitigation strategies and to create a more inclusive and equitable advertising ecosystem.
- Opting transparency – The platforms should enable the advertisers to access their bias metrics in ad delivery system to promote transparency, trust, and accountability.
- Enhancing targeting options – The platform can introduce advanced targeting parameters such as user behavioural pattern, effect of trends on user impressions and contextual relevance, along with the existing parameters. This will enhance the understanding of the advertisers in selecting their target audience.

6. CONCLUSIONS AND FUTURE DIRECTIONS

This study performed an examination of the key dimensions of AI-related advertising on the Facebook platform, including targeting, demographic representation, and geographical performance. Our findings highlight the complex interplay of targeting strategies, societal bias, and contextual factors in shaping the landscape of AI-related advertising. While interest-based targeting proves effective in reaching interested audiences, the observed gender and age disparities raise concerns about potential biases embedded within ad delivery algorithms. Furthermore, the varying advertising performance across geographical regions highlights the importance of tailored approaches that account for local market dynamics and cultural nuances. These insights have significance for both advertisers and policymakers. Overall, this study provides the following key practical implications:

- For advertisers, enhancing target precision, leveraging audience interest and

incorporating underrepresented groups will ensure better advertising strategies.

- For platform developers, addressing algorithmic biases, ensuring equitable ad distribution will mitigate the regional and gender biases.
- In terms of policies, encouraging transparency in ad delivery system will prevent and reduce unwanted social biases and ultimately improve the whole ad delivery system.

While this study provides valuable insights for AI-related advertisement strategies, some limitations must be addressed:

- Although, this study was based on Meta's robust Ad Targeting datasets, the findings may not be generalized to other platforms such as Twitter with differing target audience and engagement metrics.
- It was out of scope of this study to address the "black-box" nature of the ad-delivery algorithm, which may provide further deep understanding of how the ad-delivery system operates.
- The ad-targeting analysis can also depend on cultural and linguistic barriers which were not comprehensively addressed in this study.

Future research should address these limitations and further explore the interplay between targeting precision, demographic representation, and regional characteristics to develop more effective and inclusive AI advertising practices. Additionally, investigating the effectiveness of different messaging and creative approaches across diverse cultural contexts is crucial for ensuring inclusive and equitable representation within AI-related advertising. The role of other networking platforms such as LinkedIn on AI-related advertising should also be explored. As the influence of AI continues to evolve, it is necessary that we address the challenges and opportunities of AI-related advertisement to achieve a more transparent, responsible, and inclusive digital advertising ecosystem.

This study provided some valuable insights for advertisers seeking optimization strategies for their marketing campaigns of AI technologies. It also identified potential bias in advertisement delivery which advertisers need to consider when designing their campaigns. We examined how to reach the right audiences effectively and fairly, considering potential biases in the system. The ultimate aim of this analysis is to guide ethical and impactful advertising in this rapidly changing landscape.

7. REFERENCES

- Abuljadail, M., & Badghish, S. (2021). Exploring Type of Strategies Used by Global Brands to Engage the Saudi Consumers More in Brands' Facebook Pages in Saudi Arabia in Terms of "Like, Share and Comment". *Arab Journal of Administration*, 41(1), 405-416. <https://doi.org/10.21608/aja.2021.151257>
- Adeleye, R. A., Awonuga, K. F., Asuzu, O. F., Ndubuisi, N. L., & Tubokirifuruar, T. S. (2024). Digital marketing analytics: A review of strategies in the age of big data and AI. *World Journal of Advanced Research and Reviews*, 21(2), 073-084. <https://doi.org/10.30574/wjarr.2024.21.2.0395>
- Anderson, J., Casas Saez, G., Anderson, K., Palen, L., & Morss, R. (2019, 2019). *Incorporating context and location into social media analysis: A scalable, cloud-based approach for more powerful data science* 52nd Hawaii International Conference on System Sciences, Grand Wailea, Maui, Hawaii, USA. <http://dx.doi.org/10.24251/hicss.2019.275>
- Barbul, M., & Bojescu, I. (2023, 2023). *Generations' Perception Towards the Interaction with AI. In: R. Pamfilie, V. Dinu, C. Vasiliu, D. Pleşea, L. Tăchiciu eds* 9th BASIQ International Conference on New Trends in Sustainable Business and Consumption, Constanța, Romania. <https://doi.org/10.24818/BASIQ/2023/09/041>
- Beauvisage, T., Beuscart, J.-S., Coavoux, S., & Mellet, K. (2023). How online advertising targets consumers: The uses of categories and algorithmic tools by audience planners. *new media & society*. <https://doi.org/10.1177/14614448221146174>
- Birknerová, Z., Frankovský, M., Zbihlejšová, L., & Parová, V. (2018). Perception of advertising and expectations of advertising in terms of gender differences. *Marketing and Branding Research*, 5, 92-99. <https://doi.org/10.33844/mbr.2018.60294>
- Chander, S. (2024). Impact of artificial intelligence on society: risk and challenges. *International Journal of Engineering Science and Humanities*, 14(Special1), 103-111. <https://doi.org/10.62904/s5ezzj40>
- De Maio, C., Gallo, M., Hao, F., Loia, V., & Yang, E. (2020, 2020). *Fine-grained context-aware ad targeting on social media platforms* IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, Ontario, Canada. <http://dx.doi.org/10.1109/smc42975.2020.9282827>
- De Silva, T. M. (2023). Examining the role of advertising value perceptions in driving engagement: a framework for assessing hotel/restaurant paid advertising effectiveness on Facebook. *South Asian Journal of Business Insights*, 3(1), 03-25. <https://doi.org/10.4038/sajbi.v3i1.49>
- Dobscha, S., & Ostberg, J. (2021). Introduction to the special issue on gender impacts: consumption, markets, marketing, and marketing organisations. *Journal of Marketing Management*, 37(3-4), 181-187. <https://doi.org/10.1080/0267257x.2021.1880163>
- Fogarty, E. A. (2021). Visualizing the relationship between geographic and social media network space. *GeoJournal*, 86(6), 2483-2500. <https://doi.org/10.1007/s10708-020-10176-6>
- Gao, B., Wang, Y., Xie, H., Hu, Y., & Hu, Y. (2023). Artificial intelligence in advertising: advancements, challenges, and ethical considerations in targeting, personalization, content creation, and ad optimization. *Sage Open*, 13(4), 21582440231210759. <https://doi.org/10.1177/21582440231210759>
- Kannan, P. K. (2017). Digital marketing: A framework, review and research agenda. *International journal of research in marketing*, 34(1), 22-45. <https://doi.org/10.1016/j.ijresmar.2016.11.006>
- Kaplan, L., Gerzon, N., Mislove, A., & Sapiezynski, P. (2022, 2022). *Measurement and analysis of implied identity in ad delivery optimization* 22nd ACM Internet Measurement Conference, Nice, France. <http://dx.doi.org/10.1145/3517745.3561450>
- Kawaf, F., & Istanbuluoglu, D. (2019). Online fashion shopping paradox: The role of customer reviews and facebook marketing. *Journal of Retailing and Consumer Services*, 48, 144-153. <https://doi.org/10.1016/j.jretconser.2019.02.017>
- Kelley, P. G., Yang, Y., Heldreth, C., Moessner, C., Sedley, A., Kramm, A., Newman, D. T., & Woodruff, A. (2021, 2021). *Exciting, useful,*

- worrying, futuristic: Public perception of artificial intelligence in 8 countries AAAI/ACM Conference on AI, Ethics, and Society, <http://dx.doi.org/10.1145/3461702.3462605>
- Kolesnyk, B., & Kostynets, I. (2023). Measuring digital advertising effectiveness on facebook as part of a national brand advertising campaign. *Economics, Finance and Management Review*(3), 68-82. <https://doi.org/10.36690/2674-5208-2023-3-68-82>
- Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7), 2966-2981. <https://doi.org/10.1287/mnsc.2018.3093>
- Lin, T. H., Veeramuthu, A., Veerappan, L., Warraich, M. A., Sulistyo, P. B., & Putra, A. R. (2024). Facebook's Influence on the Effectiveness of Digital Advertising among Malaysian Youth. *Journal of Communication, Language and Culture*, 4(1), 61-83. <https://doi.org/10.33093/jclc.2024.4.1.4>
- Lulandala, E. E. (2022). Social media advertising: Concepts, contemporary trends and implications In K. Sharma & S. Aggarwal (Eds.), *Digital Marketing Outreach: The Future of Marketing Practices* (pp. 173-186). Taylor & Francis. <https://doi.org/10.4324/9781003315377-13>
- Marinchak, C. M., Forrest, E., & Hoanca, B. (2018). Artificial intelligence: Redefining marketing management and the customer experience. *International Journal of E-Entrepreneurship and Innovation (IJEI)*, 8(2), 14-24. <https://doi.org/10.4018/ijeai.2018070102>
- Meta Platforms (2024). *Meta Ad Targeting Dataset* <https://doi.org/10.48680/meta.adtargetingdataset>
- Mishra, K. D. (2017). Social media revolution -The new digital frontiers of Journalism. *Journal of advances in humanities*, 5(1), 610-620. <https://doi.org/10.24297/jah.v5i1.6068>
- Mungoli, N. (2023). Revolutionizing Industries: The Impact of Artificial Intelligence Technologies. *Journal of Electrical Electronics Engineering*, 2(3), 206-210. <https://dx.doi.org/10.33140/JEEE.02.03.03>
- Patino, A., D. Kaltcheva, V., Pitta, D., Sriram, V., & D. Winsor, R. (2014). How important are different socially responsible marketing practices? An exploratory study of gender, race, and income differences. *Journal of consumer marketing*, 31(1), 2-12. <https://doi.org/10.1108/jcm-10-2013-0733>
- Phillips, D. M., & Stanton, J. L. (2004). Age-related differences in advertising: Recall and persuasion. *Journal of Targeting, Measurement and Analysis for Marketing*, 13, 7-20. <https://doi.org/10.1057/palgrave.jt.5740128>
- Risius, M., & Ayding, O. A. I. (2018). Facebook user segmentation to enable targeted social advertisement. *39th International Conference on Information Systems, At San Francisco, USA*.
- Rodgers, W., & Nguyen, T. (2022). Advertising benefits from ethical artificial intelligence algorithmic purchase decision pathways. *Journal of business ethics*, 178(4), 1043-1061. <https://doi.org/10.1007/s10551-022-05048-7>
- Sharma, M. S., & Ashfaq, R. (2023). Targeting the digital consumer: A study of the role of social media in modern advertising. *Journal of Humanities, Music and Dance (JHMD)*, 3(04), 22-35. <https://doi.org/10.55529/jhmd.34.22.35>
- Vitalis, P. O., Onyejelem, T. E., & Okuneye, A. P. (2023). Understanding advertising in the era of social media. *Information System and Smart City*, 3(1), 502. <https://doi.org/10.59400/issc.v3i1.502>
- Vržina, A., & Pepur, M. (2021). The importance of socially responsible marketing practices on the healthy food market from the perspective of students. *Zbornik Veleučilišta u Rijeci*, 9(1), 211-226. <https://doi.org/10.31784/zvr.9.1.13>
- Whitelegge, M. (2016). Geodemographics for marketers: Using location analysis for research and marketing. *Journal of Direct, Data and Digital Marketing Practice*, 17, 258-258. <https://doi.org/10.1057/s41263-016-0008-7>
- Yu, Y. (2021). *The role and influence of artificial intelligence on advertising industry* International Conference on Social Development and Media Communication (SDMC 2021), Sanya, China. <http://dx.doi.org/10.2991/assehr.k.220105.037>

A Proposed Study of Factors Moderating Degree of Trust in LLM and ChatGPT-like Outputs

William Money
wmoney@citadel.edu

Namporn Thanetsunthorn
nthanets@citadel.edu

The Citadel Military College of South Carolina
Charleston, SC 29409

Abstract

It is important to understand the human aspect of new AI systems using Large Language Models (LLMs) like ChatGPT and Gemini in our everyday work and how they will influence the processes used to complete our activities and actions. The ways we perceive and interact with these new AI systems using generative technologies are greatly influenced by the trust we place in these technologies. This paper presents a proposed method for investigating the factors influencing the trust individuals placed in the outputs of such system tools, specifically those incorporating Large Language Models (LLMs). These systems and their components possess the capabilities to produce original work that is very highly representative of traditional human made products that reflect but do not simply duplicate the input data by predicting next-word sequences. It is well known that the system outputs may vary with respect to validity and reliability. There is a general awareness that a large language model may simply make things up. While traditional research on trust emphasizes interpersonal or interfirm trust, this proposed study will investigate the trust current and potential users place in the LLM technologies, and the factors influencing usage behaviors. Thus, the research aids in the development of a nuanced understanding of why trust is placed in AI LLM products. This understanding is crucial for designing effective tools and frameworks to introduce the tools into organizations. Trust placed in information systems has been found to be important in many domains such as business relationships, work communication, and team interactions. It shapes organizational decisions regarding systems usage. It is important to utilize AI LLM technology's capabilities and functionalities and appreciate that their use and integration into work are mediated by an individual's trust in this technology as much as one might trust in expert professional skills, and professional competence is important for a specific domain. AI LLM technology trust is therefore differentiated from trust in people to develop a deeper comprehension of users' attitudes and intentions toward this technology's adoption and usage, facilitating the development of tailored strategies and interventions in information systems research and practice.

Keywords: LLM, Trust, ChatGPT, Control, Artificial Intelligence, Credibility.

Recommended Citation 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*. v18, n4, pp 67-80. DOI# <https://doi.org/10.62273/JRRF9742>

A Proposed Study of Factors Moderating Degree of Trust in LLM and ChatGPT-like Outputs

William Money and Namporn Thanetsunthorn

1. INTRODUCTION

This paper investigates variables influencing user acceptance and belief in the outputs of ChatGPT like systems using Large Language Models (LLMs). Trust in the outputs of LLMs aids in our understanding of why users may believe and utilize the outputs. The questions broadly consider if the LLMs are viewed as being truthful, accurate, without bias, and correct, or conversely, are they seen as being highly susceptible to hallucinations and prone to creating fiction?

It builds upon the work of McKnight, Carter, Thatcher, & Clay (2011) who previously examined the role of trust within technology systems (Information Systems). They looked at the importance of trust in understanding user interactions with technology. The authors defined trust broadly as the general willingness to rely on a system or entity according to the users' perception. The perceived attributes of ability, benevolence, and integrity were used to evaluate trust in both people and technology. In considering technology trust, the 'ability' refers to perceptions of functional attributes like reliability and performance, while 'benevolence' component of technological trust relates to those that build and support the technology. Although a technology is used, users conceptualization impacts on user behavior.

The value of understanding the circumstances and perceptions associated with trust in technology when it impacts workplace interactions is important. Technology designers and implementors must recognize that users know they must rely on technology's capabilities for effective task performance. This reliance is independent of their trust in the people or the organization behind the technologies (Mcknight, Carter, Thatcher, & Clay (2011).

Mcknight, Carter, Thatcher, & Clay's (2011) work focused on the development and validation of trust technology measures. They noted that there is an important trust impact on technology adoption. It is potentially associated with a technology's acceptance and the user's post-adoption behaviors. Thus, trust influences how users employ a technology after it has been

implemented by the organization. This makes trust critical for understanding long-term usage and task and work process dependency.

The research proposed here attempts to fill a gap in the literature by focusing more directly on trust in the technology itself, rather than trust in the human or organizational entities associated with the technology. It will support the development of a more comprehensive explanation of how technological trust influences user behavior and technology acceptance. Results from prior research indicate that trust directly affects user interactions with technology, impacting everything from initial adoption to continued use. It has additional derivative impacts upon user-centric design and user support systems.

Previous research examining trust in technology has distinguished it from trust in the provider of the technology since users might trust the functionality of a software while still being skeptical about the company that produces it. Trust components like system reliability, user support, and perceived utility play critical roles in forming trust in technology itself.

Initial user trust in technology and systems is modeled and analyzed by Li, Hess, Valacich, (2008). They viewed the initial trust as being crucial for overcoming users' initial perceptions of risk and uncertainty when adopting new technologies. Their research suggests that initial trust forms because users must rely on secondary information and preconceived expectations about a technology's characteristics before actual use. Indirect information such as the technology's perceived attributes, strong organizational backing, and societal endorsements shape and form the users' attitudes and subsequent decisions to adopt and trust in the technology.

Personality factors, cognitive assessments of the technology's reliability and effectiveness, calculative judgments on the benefits versus risks, and institutional factors prompt initial trust assessments (Li, Hess, & Valacich, 2008). These elements collectively contribute to initial trusting beliefs and intentions. Organizations that seek to build positive user first impressions and encourage technology gain value from understanding how these first impressions are

formed. They may then improve adoption, and more effectively direct initial user perceptions.

The impact of interacting with ChatGPT, a language model developed by OpenAI, has been assessed previously by examining its relationship with trust, user perception, stereotype perception, and two psychological outcomes: self-esteem and psychological well-being (Salah, Alhalbusi, Ismail, & Abdelfattah, 2023). The research study hypothesizes that there is a positive direct relationship between trust in ChatGPT, user perception, and stereotype perception of ChatGPT with self-esteem. Job anxiety was also hypothesized to be a moderator of the relationship between user perception of ChatGPT and psychological well-being. Stereotyped perceptions of ChatGPT were found to significantly predict self-esteem, while user perception and trust in ChatGPT had a positive direct relationship with self-esteem based on this work. Job anxiety moderates the relationship between user perception of ChatGPT and psychological well-being. The hypothesized psychological effects of AI technology are supported by these data.

Users have reason to mistrust generative models according to research on these tools. Their tendency to “hallucinate” or make up responses and generate outputs that are biased or may contain harmful content has been described in many publications and blogs. Schulman, Zoph, Kim, Hilton, Menick, Weng, J., ... & Ryder (2022) trained a ChatGPT model and described a number of potential problems with the output. These included: ChatGPT sometimes writes apparently plausible-sounding but incorrect or nonsensical answers; declines responding to questions that it could answer correctly; ChatGPT responses are sensitive to tweaks to the input phrasing or attempting the same prompt multiple times with is answering correctly or incorrectly based on prompt variations; it often excessively verbose and overuses certain phrases and over-optimization; does not ask clarifying question and usually guesses what the user intended; and will sometimes respond to harmful instructions or exhibit biased behavior.

For example, Alkaissi, & McFarlane (2023) instructed ChatGPT to write about the pathogenesis of two conditions - homocystinuria-associated osteoporosis, and a rare metabolic disorder, late-onset Pompe disease (LOPD). The results found negative aspects of the chatbot's performance. Comparing it to the US Medical Licensing Examination (USMLE) Step 1, Step 2 CK, and Step 3, as open-ended and multiple-

choice questions (MCQ). The result showed the accuracy was low indicating that the performance is tied to perception and understanding of the subject. The authors note that the written outputs are credible, but that generated data mixes true and completely fabricated data,

2. RELATED WORK

Trust

The literature on trust and automation systems suggests that LLMs may be subject to moderation by several variables that would both promote or discourage trust and therefore influence the user's action regarding the outputs of an LLM. The user might have a tendency to disregard the model's response or question the outputs of these systems or conversely accept the results without checking the facts against known values or original reputable sources.

The issues with trust in LLM and ChatGPT like output is significant for many reasons (Brzowski, & Nathan-Roberts, November 2019). The authors argue that a lack of human users' trust is due to the limited semantic understanding between humans and similar systems. They posit that the communication between the user and the LLMs, such as ChatGPT, may be used to develop greater degrees of trust because they offer an interactive collaboration approach. The authors assessed the impact of ChatGPT on trust in a human-robot collaboration assembly task. A robot control system used ChatGPT to control a 7-degree-of-freedom robot arm. The arm retrieved and placed tools using natural language control issues by humans. The user's trust measured by attitude surveys was increased. This was attributed to the Chatbot understanding the nuances of human language and responding appropriately. The findings of this study suggest that the development of trust can be improved after experience and with positive results.

The value of trust in technology and especially new technologies such as LLMs has long been a topic of study in the information systems literature. Trust has been examined in the information systems domain. It has been shown to be important in explaining the adoption and use of new technologies such as the usage of systems in e-commerce, and virtual communities (Söllner, & Leimeister, 2013). These authors examined a body of knowledge on trust regarding its reliability and the antecedents of trust in the information systems literature. They examined many different antecedents for different trust relationships in different contexts. They found that measurement model mis-specification issues

could be serious challenges in information systems trust research. The most common issue involved using formative indicators in reflective measurement models. This could threaten the strength of the association found in the structural relationships between trust and its antecedents in these studies.

Lowry et al. (2008) and Vance et al. (2008) research addresses measurement model mis-specification and the use of second-order measurement models to assess the trust in systems. These researchers report that the work by (Klein & Rai, 2009; Venkatesh & Bala, 2012) was valuable and solid. Klein, & Rai (2009) found that trust was very important as an aid in strategic information flows between buyers and suppliers within logistics supply chain relationships. It positively impacted other relationship-specific performance outcomes. Trust results in the valuable development of cooperative initiatives and relationships rather than conventional “arms-length” transactional exchanges. The partnerships are not limited to the sharing of order-related information and extend to strategic information that has value for both parties. The Venkatesh, & Bala (2012) research on the inter-organizational business process standards (IBPS) found the standards are adopted because of trust factors that represent synergies between a focal firm and its trading partners. Their study of 248 firms (124 dyads) in the high-tech industry also found that relational trust had direct effects on IBPS adoption.

Salah, Alhalbusi, Ismail, & Abdelfattah (2023) investigated generative AI tool adoption (ChatGPT and Bard) in public administration and street-level bureaucracy. They identify several benefits from the use of these powerful tools including insights into bureaucratic behavior and decision-making processes, and citizen interactions. However, they also recognize that the complex nature of AI algorithms (such as those applied by ChatGPT) poses difficulties for researchers' and stakeholders' comprehension of the decision-making processes behind AI-generated insights. Concerns about accountability and trust in AI-driven research findings may result from this lack of algorithmic transparency. They recommend that clear explanations of the AI algorithms and their implications be provided with the outputs.

Self Efficacy

Self-efficacy has been shown to be associated with and an influencer of trust in a variety of commerce and technology situations. Trust has been recognized as a critical factor for electronic

commerce because online transactions are characterized as a process that involves uncertainty and risk. Achieving a high degree of trust is an effective means of reducing uncertainty and risk. Kim, & Kim (2005, January) research describe self-efficacy as having an impact on trust building and uncertainty reduction. The results show that self-efficacy affects trust in the web vendor and positively influences purchase intentions.

Abdunabi, Hbaci, Center, & Nyambe (2023) examined perceived programming self-efficacy of information system students as a factor helping students learn to program. Their examination of students' internal characteristics and programming self-efficacy found a strong connection. Their survey assessed students' beliefs in their programming competence, value attributed to learning programming, time spent practicing, and instructional guidance frequency. The value students placed on learning programming was described as the most significant variable associated with programming self-efficacy.

Internet banking (IB) has also been investigated as an outcome impacted by four factors - hedonic motivation, habit, self-efficacy and trust using a survey questionnaire that collected data for structural equation modelling (SEM). These research findings strongly supported the conceptual model by explaining 73% of variance in behavioral intention to use internet banking (Alalwan, Dwivedi, Rana, Lal, Williams, 2015). Further, hedonic motivation, habit, self-efficacy and trust are all confirmed to have significant influences on behavioral intention. Trust was found to be profoundly predicted by both self-efficacy and hedonic motivation.

Chamorro-Koc, Peake, Meek, & Manimont (2021) researched the growing commercial market for wearable health technology. But they value is questioned by their work due to the lack of validation and abandonment rates. Self-efficacy mechanisms are being incorporated into the design of health technologies, through (i) past experience, (ii) tracking of activities , (iii) autonomy, (iv) strong interest in personal health, and (v) reliability and validity of data impacts on confidence in health technologies. Their conceptual model offers support for improving self-efficacy and trust in health technologies so designers and developers can incorporate these factors into design features for effective personal health technology.

Perceived Control. Humans and intelligent agent interactions are very important in today's world because of the large number of services and controls that are available to individual management. Research on human agent interaction (HAI) has therefore become important since effectively controlling the agents can improve efficiency and interactions. Liao, Li, Cheng, & Yang (2023) assert that at some point human will have negative emotions (toward agents) such as panic, fear, and disgust of the very effective. The study defines perceived control as the degree of confidence people have in interacting with intelligent agents. It is seen as an overall evaluation and attitude of intelligent agents' feeling of control. Thus, high perceived control of intelligent agents is a good description of a desired human relationship with HAI. Perceived control represents a sense of internal control based on the ability, knowledge, skills, or familiarity that produces cognitive and decisional control.

Technology Acceptance Model (TAM). Decisions regarding the acceptance or rejection of new technology have open question as new systems and technologies have had greater and greater impacts upon people's lives and work environments. The reasons behind acceptance and the factors that influence acceptance have been assessed with the technology acceptance model (TAM) for approximately 35 years. The model stems from the psychological theory of reasoned action and theory of planned behavior. It has aided greatly in our understanding of the predictors of human behavior toward prospective acceptance or rejection of a technology. The model has been extended and modified to apply to a variety of information systems and related technologies. The body of research has revealed new factors that can significantly influence the TAM core variables Holden, & Karsh (2010). TAM is understood to contain six causally related constructs: perceived ease of use, perceived usefulness, attitude towards using, behavioral intention to use and actual system use (Davis, Bagozzi & Warshaw, 1989; Erasmus, Rothmann, & Van Eeden, 2015).

Trust has been found to be an important concept that can be integrated with TAM. For example, Pavlou's (2003) research applied the TAM model variables (perceived usefulness and ease of use) to a technology-driven environment to predict e-commerce acceptance. Pavlou integrated trust and perceived risk (uncertainty of the environment) with TAM. The research findings strongly support the proposed model, showing that trust was an indirect antecedent acting

through risk perception. Additionally, research by Wu, Zhao, Zhu, Tan, & Zheng (2011) identified trust as an important factor that influences the user's online behavior. This role of trust on subject type (students or non-students) and context type (commercial or non-commercial) significantly influenced TAM constructs.

3. RELEVANT CONTROL VARIABLE

Demographic Factors.

Trust in e-vendors and their technologies implemented through IT and Web site interfaces is a multifaceted construct influenced by various factors (Gefen et al., 2003). Building upon previous research, scholars have explored and identified numerous factors as significant predictors of individuals' propensity to trust in technologies (including systems like ChatGPT). Thus, it is essential to consider these variables as control variables when conducting surveys to measure the level of trust in technologies. This approach can effectively isolate the potential influence of specific factors, thereby yielding a more accurate understanding of users' attitudes regarding trust in technologies. Notably, demographic factors and individual differences in personality traits emerge prominently among the factors contributing to trust in technologies (e.g., Choung et al., 2023; McElroy et al., 2007; Sundar, 2020; Svendsen et al., 2013; Venkatesh et al., 2003).

Regarding demographic variables such as age, gender, level of education, and socioeconomic status, there is a general consensus among researchers that including these variables in surveys allows for a better understanding of how trust in technology varies across different demographic groups and population segments (Gefen et al., 2003; Venkatesh et al., 2003). In particular, previous research examining technology acceptance models has documented that age plays a crucial role in how people adopt technologies and trust automation (e.g., Hoff & Bashir, 2015; Morris & Venkatesh, 2000). For example, older individuals tend to prefer human editors over balancing algorithms for news story consumption (Thurman et al., 2019). They also tend to be more skeptical than younger people about the fairness of decisions made by automation, robots and AI (Hoff & Bashir, 2015; Oksanen et al., 2020). This difference may be attributed to varying levels of familiarity and comfort with technology, with younger individuals, who are more exposed to and familiar with technology, showing higher levels of trust (Morris & Venkatesh, 2000).

There have been scholarly efforts dedicated to investigating whether gender is a significant predictor of the use of AI tools and how perceptions of AI tools vary by gender. Previous research consistently shows that gender influences how individuals interact with AI technologies. For example, women are often perceived as underrepresented in the fields of technology with a study of social robot use (De Graaf & Allouch, 2013). They are also shown to be under-represented as users and creators in using AI-based tools in a STEM study of women. The study found they are thereby limited (by gender) in their access to and utilization of AI tools (Ofosu-Ampong, 2023). Gender differences can also reveal varying perceptions and attitudes toward new technology (Venkatesh & Davis, 1996; Venkatesh & Morris, 2000). In their seminal work, Venkatesh and Morris (2000) conducted a five-month survey involving 342 workers regarding the transition to a new software system. The survey results indicate that men tend to base their technology usage decisions more heavily on perceived usefulness compared to women. Conversely, women are more influenced by perceptions of ease of use and social norms.

In addition to age and gender, levels of education and socioeconomic status are widely recognized as significant factors influencing the level of trust individuals place in technologies. Previous research suggests that higher levels of educational attainment are often linked to greater critical thinking skills and a better understanding of complex technologies, leading to more informed and nuanced trust in social networking sites (Hargittai & Hsieh, 2010), Internet usage types (Van Deursen & Van Dijk, 2014), and AI in medicine for radiology, robotic surgery, and dermatology (Yakar et al., 2022). Specifically, individuals with higher education levels are more likely to utilize AI technologies and make informed judgments about their reliability and benefits. Similarly, socioeconomic status can influence trust in AI by affecting access to technology and related resources. Individuals with higher incomes often have greater exposure to and familiarity with advanced technologies, which can cultivate a more trusting attitude toward AI (Van Deursen & Van Dijk, 2014; Zhang & Dafoe, 2019). These individuals are also more likely to experience the benefits of AI in their daily lives, subsequently reinforcing their trust in AI technologies. On the other hand, those with lower socioeconomic status may have limited access to technology, leading to less familiarity and potentially more skepticism about AI technologies. The significance of education level

and socioeconomic status in shaping perceptions and acceptance of AI technologies is further highlighted in the work of Choung et al. (2023). Their survey of 525 respondents from the general U.S. population demonstrates that adults with higher levels of education and income tend to exhibit greater trust in AI.

Personality Traits

Human-related factors beyond demographics are widely recognized as critical determinants of individuals' technology trust and Internet use (McElroy et al., 2007), human-AI interaction (Sundar, 2020), and consumer use of technology (Venkatesh, Thong, & Xu, 2012). This body of literature predominantly focuses on the Five-Factor Model of personality traits, commonly known as the Big Five, which encompasses agreeableness, openness, conscientiousness, extraversion, and neuroticism (Digman, 1990; John et al., 2008). The model has been a focal point in the existing literature for evaluating how personality traits may influence individuals' willingness to trust in technologies. Numerous studies utilizing the Big Five have demonstrated that these traits can significantly impact individuals' trust in technologies, underscoring the importance of considering personality when developing designs for technologies and when implementing systems. Below, we discuss some notable studies in this area.

The majority of previous studies indicate a positive correlation between agreeableness and trust in human-centered AI interfaces (Böckle et al., 2021), technology acceptance (Devaraj et al., 2008), and trust in automated vehicles (Kraus et al., 2020). In their influential work, Park and Woo (2022) investigated affective and cognitive attitudes toward AI. They found that individuals with high agreeableness scores tend to hold positive attitudes toward AI, particularly regarding its perceived usefulness. Similarly, consistent research findings indicate that individuals with high levels of openness tend to exhibit favorable attitudes toward AI. For example, Antes et al. (2021) conducted research on attitudes toward AI driven healthcare technologies, and Oksanen et al. (2020) have reported evidence from an online AI trust game that openness to experience is strongly correlated with greater trust in AI systems. Their work supports a previous DeYoung et al. (2007) finding that individuals with high levels of openness are more likely to seek out new information and experiences. This propensity for exploration and curiosity likely contributes to individuals' higher levels of trust and acceptance of new technologies (McElroy et al., 2007; Svendsen et al., 2013).

The literature also indicates that extraversion and conscientiousness play significant roles in shaping individuals' trust in machine characteristics and auto use (Merritt & Ilgen, 2008), AI based voice technologies (Bawack et al., 2021), and in AI voice shopping (Kraus et al., 2020). Extraverts, characterized by their sociability and enthusiasm, are more likely to adopt AI-driven systems, such as robots and virtual assistants, due to their preference for social interaction (Kaplan et al., 2019; Oksanen et al., 2020). Similarly, conscientiousness, which reflects traits such as diligence and carefulness, has been found to correlate positively with trust in cloud customer relationship management technology by Fu, & Chang (2016). This finding support the position that conscientious individuals tend to value the reliability and efficiency of information systems, resulting in higher levels of trust in these technologies. McKnight et al. (2002) further argue that the methodical and organized nature of conscientiousness aligns well with the structured and predictable aspects of information systems. This alignment implies that conscientious individuals are more likely to trust technology due to their propensity to appreciate the reliability and consistency that information systems offer. On the other hand, individuals with lower levels of neuroticism, which indicates emotional stability, tend to be more accepting of technology. Prior studies show that individuals scoring low on neuroticism tend to experience less anxiety and distrust, leading to a more positive attitude toward AI technologies (Kraus et al., 2020; Sharan & Romano, 2020, Zhang et al., 2020). This reduced anxiety enables them to engage more confidently with AI systems, thereby enhancing their trust in such technologies.

4. CONCEPTUAL FRAMWORK

Our research framework is based upon the psychological theories of reasoned action and theory of planned behavior as is the TAM body of research. We seek to expand our understanding of the role of trust from the perspective of the individual, and our appreciation of the important role that predictors of human trust in LLM and AI technology. (Holden,& Karsh (2010); Davis, Bagozzi & Warshaw, 1989; Erasmus,, Rothmann, & Van Eeden, 2015).

5. METHODOLOGY

We developed the following 5 primary hypotheses for this study based in the trust literatures suggesting propensities to trust in the AI LLM technology, and the related literature.

- H1. Perception of High self-efficacy will positively impact the level of Trust in LLMs.
- H2. Perception of High-Control will positively impact the level of Trust in LLMs.
- H3. Perception of High-Usefulness will positively impact the level of Trust in LLMs.
- H4. Perception of High Ease of Use will positively impact the level of Trust in LLMs.
- H5. High Intention to Use will positively impact the level of Trust in LLMs.
- H6. Control Variable will show significant differences in intention to use and use of AI ChatGPT technologies among sub-populations.

Data for this research will be collected with a survey questionnaire administered to graduate and undergraduate students in the summer and fall semesters, 2024. (The number of participants will depend upon enrollment and sections participation.) It is important to note that the researchers expect the graduate and undergraduate classes to have significant difference when categorized by the control variables. The graduate students are primarily part-time and employed. The undergraduates are younger (compared to the graduates), full time, unemployed, and with little or no earned income. The respondents' demographics (ages ranges, sex, education levels, etc.) will be reported and used in the analyses.

The students will be asked to offer response with and about their trust and their use of using an LLM or ChatGPT like system. Students will be provided a link to the survey questionnaire randomly distributed using MS Forms.

SPSS application (Version-20) or SAS 9.4 was used to analyze the data. The instrument used for this study was designed based on the focus of trust, the investigation objective of the study. The reliability and validity of the instruments will be calculated and reported.

Survey data were collected using a five-point Likert scale (1 for strongly disagree to 5 strongly agree). The survey question are adapted from existing survey scales from prior research. The survey guidance will state that the questionnaire investigates students' opinions about their trust in the use of ChatGPT and other LLMs.

The survey questions are show in the appendix.

6. DISCUSSION

We recognized there will be several important limitations to this work. First, this study only addresses generative AI LLMs, and only one

specific tool (ChatGPT) will be referred to in the survey questionnaire. Thus, the results may not be widely transferable, and other forms of generative technology (RAG -Research Augmented Generation), and other tools that may be used by the respondents. Secondly, the trust measures may have different meanings for different populations. Trust, due to one's inherent belief in technology, may vary based on the task performed and the situation or context of the work. The student sample used to collect the data may not represent a more general population and may not address the context and nuances of the situations where AI and GPT is eventually employed. Finally, the student population may not effectively represent the organization member who is to use and apply AI in a work environment.

Unfortunately, we have no hard measures to compare our result with actual access and use of AI and ChatGPT in producing work products. We believe would be informative to know if individuals are actually using the LLMs, and the extent of the usage and reliance on these products.

7. CONCLUSIONS

Our conclusion will depend upon the study results and detailed analysis of trust and the control variables. However, we believe there is no question that AI and Chat like LLMs may add great value and save user time for some tasks. They are and will be used by organizations and the public to for work productivity improvements. We hope to help answer important questions - who will place trust in the output of these tools and use them in important or valued work? Does trust in AI and specifically ChatGPT like products compare favorably with existing models describing continued postadoption of its use. Significant questions for additional research will exist after our work. For example, does the influence of trust in this new AI vary over time? Will belief in technology improve as the products mature and evolve to provide new features, and how will product evolution take to impact adoption behavior? Finally, future work may help to determine if trust in AI may mediate the influence of trust in people who promote, develop, or support a specific AI product. Conversely, it is not clear if trust in AI and ChatGPT like successes can influence trust in people to build or deploy the technology? Our future research will explore these questions.

8. RECOMMENDATIONS

Our recommendation for the users, designers, developers and LLM technologies will be based upon our findings and discussion of the issues associated with this new technology..

9. REFERENCES

- Abdunabi, R., Hbaci, I., Center, C. S., & Nyambe, T. (2023). Predicting Perceived Programming Self-Efficacy for Information System Students. In Proceedings of the ISCAP Conference ISSN (Vol. 2473, p. 4901).
- Alalwan, A. A., Dwivedi, Y. K., Rana, N. P., Lal, B., & Williams, M. D. (2015). Consumer adoption of Internet banking in Jordan: Examining the role of hedonic motivation, habit, self-efficacy and trust. *Journal of Financial Services Marketing*, 20, 145-157.
- Alkaiissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, 15(2).
- Antes, A. L., Burrous, S., Sisk, B. A., Schuelke, M. J., Keune, J. D., & DuBois, J. M. (2021). Exploring perceptions of healthcare technologies enabled by artificial intelligence: an online, scenario-based survey. *BMC Medical Informatics and Decision Making*, 21(1), 221. <https://doi.org/10.1186/s12911-021-01586-8>
- Bawack, R. E., Wamba, S. F., & Carillo, K. D. A. (2021). Exploring the role of personality, trust, and privacy in customer experience performance during voice shopping: Evidence from SEM and fuzzy set qualitative comparative analysis. *International Journal of Information Management*, 58, 102309. <https://doi.org/10.1016/j.ijinfomgt.2021.102309>
- Böckle, M., Yeboah-Antwi, K., & Kouris, I. (2021). Can you trust the black box? the effect of personality traits on trust in AI-enabled user interfaces. In Degen, H., & Ntoa, S. (Ed.), *Artificial Intelligence in HCI. HCII 2021. Lecture notes in computer science* (Vol. 12797, pp. 3-20). Springer. https://doi.org/10.1007/978-3-030-77772-2_1
- Brzowski, M., & Nathan-Roberts, D. (2019, November). Trust measurement in human-automation interaction: A systematic review. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 63, No. 1, pp. 1595-1599). Sage CA: Los

- Angeles, CA: SAGE Publications.
- Chamorro-Koc, M., Peake, J., Meek, A., & Manimont, G. (2021). Self-efficacy and trust in consumers' use of health-technologies devices for sports. *Heliyon*, 7(8).
- Choung, H., David, P., & Ross, A. (2023). Trust and ethics in AI. *AI & SOCIETY*, 38(2), 733-745.
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management science*, 35(8), 982-1003.
- De Graaf, M. M., & Allouch, S. B. (2013). Exploring influencing variables for the acceptance of social robots. *Robotics and Autonomous Systems*, 61(12), 1476-1486.
- Devaraj, S., Easley, R. F., & Crant, J. M. (2008). How does personality matter? Relating the Five-Factor Model to technology acceptance and use. *Information Systems Research*, 19(1), 93-105.
- DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, 93(5), 880-896.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1), 417-440.
- Erasmus, E., Rothmann, S., & Van Eeden, C. (2015). A structural model of technology acceptance. *SA Journal of Industrial Psychology*, 41(1), 1-12.
- Fu, H. P., & Chang, T. S. (2016). An analysis of the factors affecting the adoption of cloud consumer relationship management in the machinery industry in Taiwan. *Information Development*, 32(5), 1741-1756.
- Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly*, 27(1), 51-90.
- Hargittai, E., & Hsieh, Y. P. (2010). Predictors and consequences of differentiated practices on social network sites. *Information, Communication & Society*, 13(4), 515-536.
- Hargittai, E., & Hsieh, Y. P. (2010). Predictors and consequences of differentiated practices on social network sites. *Information, Communication & Society*, 13(4), 515-536.
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407-434.
- Holden, R. J., & Karsh, B. T. (2010). The technology acceptance model: its past and its future in health care. *Journal of biomedical informatics*, 43(1), 159-172.
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and Research*, 3(2), 114-158.
- Kaplan, A. D., Sanders, T., & Hancock, P. A. (2019). The relationship between extroversion and the tendency to anthropomorphize robots: A Bayesian analysis. *Frontiers in Robotics and AI*, 5, 135.
- Kim, Y. H., & Kim, D. J. (2005, January). A study of online transaction self-efficacy, consumer trust, and uncertainty reduction in electronic commerce transaction. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (pp. 170c-170c). IEEE.
- Klein, Richard & Rai, Arun. (2009). Interfirm Strategic Information Flows in Logistics Supply Chain Relationships. *MIS Quarterly*, 33(4), 735-762.
- Kraus, J., Scholz, D., & Baumann, M. (2020). What's driving me? Exploration and validation of a hierarchical personality model for trust in automated driving. *Human Factors*, 63(6), 1076-1105.
- Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems*, 17(1), 39-71.
- Liao, X., Li, X., Cheng, Z., & Yang, Y. (2023). *Perceived control in human-agent interaction: Scale development and validation*. 19(02), 597.
- Madsen, M., & Gregor, S. (2000, December). Measuring human-computer trust. In *11th Australasian conference on information systems* (Vol. 53, pp. 6-8).
- McElroy, J. C., Hendrickson, A. R., Townsend, A. M., & DeMarie, S. M. (2007). Dispositional factors in Internet use: Personality versus cognitive style. *MIS Quarterly*, 31(4), 809-820.
- Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information*

- systems (TMIS), 2(2), 1-25.
- McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3), 334-359.
- Morris, M. G., & Venkatesh, V. (2000). Age differences in technology adoption decisions: Implications for a changing work force. *Personnel Psychology*, 53(2), 375-403.
- Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors*, 50(2), 194-210.
- Oforu-Ampong, K. (2023). Gender Differences in Perception of Artificial Intelligence-Based Tools. *Journal of Digital Art & Humanities*, 4(2), 52-56.
- Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human-technology interactions online. *Frontiers in Psychology*, 11, 568256. <https://doi.org/10.3389/fpsyg.2020.568256>
- Park, J., & Woo, S. E. (2022). Who likes artificial intelligence? Personality predictors of attitudes toward artificial intelligence. *The Journal of Psychology*, 156(1), 68-94.
- Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International journal of electronic commerce*, 7(3), 101-134.
- Salah, M., Alhalbusi, H., Ismail, M. M., & Abdelfattah, F. (2023). Chatting with ChatGPT: decoding the mind of Chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being. *Current Psychology*, 1-16.
- Schulman, J., Zoph, B., Kim, C., Hilton, J., Menick, J., Weng, J., ... & Ryder, N. (2022). Chatgpt: Optimizing language models for dialogue. OpenAI blog, 2, 4.
- Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. J. Weinman, S. Wright, & M. Johnston, Measures in health psychology: A user's portfolio. Causal and control beliefs, 35(37), 82-003.
- Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, 6(8), e04572. <https://doi.org/10.1016/j.heliyon.2020.e04572>
- Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human-AI interaction (HAI). *Journal of Computer-Mediated Communication*, 25(1), 74-88.
- Söllner, M., & Leimeister, J. M. (2013). What we really know about antecedents of trust: A critical review of the empirical information systems literature on trust. *Psychology of Trust: New Research*, D. Gefen, Verlag/Publisher: Nova Science Publishers.
- Svendsen, G. B., Johnsen, J.-A. K., Almås-Sørensen, L., & Vittersø, J. (2013). Personality and technology acceptance: The influence of personality factors on the core constructs of the Technology Acceptance Model. *Behaviour & Information Technology*, 32(4), 323-334.
- Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2019). My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism*, 7(4), 447-469.
- Van Deursen, A. J., & Van Dijk, J. A. (2014). The digital divide shifts to differences in usage. *New Media & Society*, 16(3), 507-526.
- Venkatesh, V. & Bala, H. (2012). Adoption and Impacts of Interorganizational Business Process Standards: Role of Partnering Synergy. *Information Systems Research*, 23(4), 1131-1157. doi: 10.1287/isre.1110.0404
- Venkatesh, V., & Davis, F. D. (1996). A model of the antecedents of perceived ease of use: Development and test. *Decision Sciences*, 27(3), 451-481.
- Venkatesh, V., & Morris, M. G. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. *MIS Quarterly*, 24(1), 115-139.
- 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.
- Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly*, 36 (1), 157-178.

- Wu, K., Zhao, Y., Zhu, Q., Tan, X., & Zheng, H. (2011). A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type. *International Journal of Information Management*, 31(6), 572-581.
- Yakar, D., Ongena, Y. P., Kwee, T. C., & Haan, M. (2022). Do people favor artificial intelligence over physicians? A survey among the general population and their view on artificial intelligence in medicine. *Value in Health*, 25(3), 374-381.
- Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly*, 36 (1),157-178.
- Zhang, B., & Dafoe, A. (2019). Artificial intelligence: American attitudes and trends. Available at SSRN 3312874.
- Zhang, T., Tao, D., Qu, X., Zhang, X., Zeng, J., Zhu, H., & Zhu, H. (2020). Automated vehicle acceptance in China: Social influence and initial trust are key determinants. *Transportation Research Part C: Emerging Technologies*, 112, 220–233. <https://doi.org/10.1016/j.trc.2020.01.027>

Appendix 1 (Qualtrics Survey)

Survey questions are provided in this appendix.

These questions have been carefully reviewed and rewarded as appropriate by replacing the "system" terminology with "LLM or chat GPT like" system designation for clarity in the appropriate questions. The survey is currently complete in Qualtrics.

Survey Introduction

Welcome to our new technology survey!

This survey aims to gather feedback and insights on the user experience and perceptions of the Large Language Model (LLM) or ChatGPT like system. Even if you have not tried or used an LLM, we would like you to share your thoughts about this new technology.

The LLM is an advanced language model that uses artificial intelligence (AI) technology to generate human-like text responses based on users' queries or prompts. It can engage in conversations, answer questions, provide explanations, and generate creative content across a wide range of topics. ChatGPT, created by OpenAI, is one example of the LLM.

Your feedback is valuable in helping us understand how users interact with and perceive this cutting-edge technology. Your responses will always remain anonymous and confidential.

Thank you for taking the time to participate in our survey!

Demographic Questions (Select appropriate category)

Age, Gender, Level of Education, Employment Status (student – not employed. Student -employed - FT), Household Income (and N/A), Ethnicity/Race, Marital Status, Number of Children

Use and Knowledge of LLMs (Like ChatGPT)

- I have never Hear of this technology.
- I have heard or read about it but have not used it yet.
- tried it once or twice – free versions.
- I use it infrequently (every 3-4 months).
- I use it monthly.
- I use it weekly.
- I use it daily.
- I use it very often each day.
- I have purchased a subscription and pay for its use. (Yes, No.)

Likert Scale:

1 = Strongly agree, 2 = Agree, 3 = Neutral, 4 = Disagree, 5 – Strongly agree

Personality

Personality (I see myself as:)

- Someone who is reserved.
- Someone who is generally trusting.
- Someone who tends to be lazy.
- Someone who is relaxed, handles stress well.
- Someone who has few artistic interests.
- Someone who is outgoing, sociable
- Someone who tends to find fault with others.
- Someone who does a thorough job.
- Someone who gets nervous easily.
- Someone who has an active imagination.

Trust questions (Madsen, & Gregor, 2000)

- R1 - The system always provides the advice I require to make my decision.
- R2 - The system performs reliably.
- R3 - The system responds the same way under the same conditions at different times.
- R4 - I can rely on the system to function properly.
- R5 - The system analyzes problems consistently.

2. Perceived Technical Competence

- T1 - The system uses appropriate methods to reach decisions.
- T2 - The system has sound knowledge about this type of problem built into it.
- T3 - The advice the system produces is as good as that which a highly competent person could produce.
- T4 - The system correctly uses the information I enter.
- T5 - The system makes use of all the knowledge and information available to it to produce its solution to the problem.

3. Perceived Understandability

- U1 - I know what will happen the next time I use the system because I understand how it behaves.
- U2 - I understand how the system will assist me with decisions I have to make.
- U3 - Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.
- U4 - It is easy to follow what the system does.
- U5 - I recognize what I should do to get the advice I need from the system the next time I use it.

4. Faith

- F1 - I believe advice from the system even when I don't know for certain that it is correct.
- F2 - When I am uncertain about a decision I believe the system rather than myself.
- F3 - If I am not sure about a decision, I have faith that the system will provide the best solution.
- F4 - When the system gives unusual advice I am confident that the advice is correct.
- F5 - Even if I have no reason to expect the system will be able to solve a difficult problem, I still feel certain that it will.

5. Personal Attachment

- P1 - I would feel a sense of loss if the system was unavailable and I could no longer use it.
- P2 - I feel a sense of attachment to using the system.
- P3 - I find the system suitable to my style of decision making.
- P4 - I like using the system for decision making.
- P5 - I have a personal preference for making decisions with the system.

Perceived Control Questions

Perceived Control. Liao, Li, Cheng, & Yang (2023) The scale includes affective control, cognitive control

and conative control .

Affective control

- (F1) AC1 The intelligent agent is always trying to entertain me.
- AC2 The intelligent agent is very polite to me.
- AC3 The intelligent agent only cares about me.
- AC4 The intelligent agent does not get angry.
- AC5 The intelligent agent makes me feel superior.

Cognitive control

- CgC1 Human beings dominate the intelligent agent.
- CgC2 I understand how the intelligent agent works well.
- CgC3 The intelligent agents are designed to serve us.
- CgC4 I know how to use intelligent agents very well.

Conative control

- CaC1 I can dictate the behavior of intelligent agents.
- CaC2 The intelligent agent obeys me.
- CaC3 The intelligent agent only acts when I allow it.

Self-Efficacy (Schwarzer, 1992).

- 1 I can always manage to solve difficult problems if I try hard enough.
- 2 If someone opposes me, I can find the means and ways to get what I want.
- 3 It is easy for me to stick to my aims and accomplish my goals.
- 4 I am confident that I could deal efficiently with unexpected events.
- 5 Thanks to my resourcefulness, I know how to handle unforeseen situations.
- 6 I can solve most problems if I invest the necessary effort.
- 7 I can remain calm when facing difficulties because I can rely on my coping abilities.
- 8 When I am confronted with a problem, I can usually find several solutions.
- 9 If I am in trouble, I can usually think of a solution.
- 10 I can usually handle whatever comes my way.

Namporn - TAM Questions (modified from Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 13(3), 319-340.)

Perceived Usefulness:

- 1. Using AI technologies enable me to accomplish tasks more quickly.
- 2. Using AI technologies improve my job performance.
- 3. Using AI technologies increase my productivity.
- 4. Using AI technologies enhance my effectiveness on the job.
- 5. Using AI technologies make it easier to do my job.
- 6. I find AI technologies useful in my job.

Perceived Ease of Use:

- 1. It is easy for me to learn how to use AI technologies.
- 2. It is easy for me to make AI technologies do what I want them to do.
- 3. I find my interaction with AI technologies is clear and understandable.
- 4. I find AI technologies to be flexible to interact with.
- 5. It is easy for me to become skilled in using AI technologies.
- 6. I find AI technologies easy to use.