

# **Bridging Technology and Therapy: Assessing the Quality and Analyzing the Impact of Human Editing on AI-Generated SOAP Notes in Pediatric Rehabilitation**

by

Solomon Amenyo

A thesis  
presented to the University of Waterloo  
in fulfillment of the  
thesis requirement for the degree of  
Master of Mathematics  
in  
Computer Science

Waterloo, Ontario, Canada, 2025

© Solomon Amenyo 2025

### **Author's Declaration**

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Abstract

This thesis explores the integration of artificial intelligence (AI) into clinical documentation, focusing on evaluating AI-generated SOAP (Subjective, Objective, Assessment, and Plan) notes in pediatric rehabilitation settings. AI-powered tools, such as Microsoft's Copilot and the University of Waterloo-developed KAUWbot, offer potential efficiencies by automating aspects of clinical documentation. However, their quality, reliability, and applicability to clinical practice have remained largely unexamined.

The research aims to assess and compare the quality of human-authored SOAP notes and AI-generated notes across five key criteria: *clarity*, *completeness*, *conciseness*, *relevance*, and *organization*. A dataset of 432 SOAP notes, divided into four pools, was evaluated using a custom rubric. The pools included human-authored notes, Copilot-generated notes edited by occupational therapists, unedited KAUWbot-generated notes, and KAUWbot-generated notes edited by occupational therapists. A rigorous anonymization process ensured evaluator impartiality.

Findings indicate that AI-generated notes, particularly when edited by clinicians, achieve comparable or superior quality to human-authored notes. After editing, notes generated by KAUWbot, a model fine-tuned on pediatric occupational therapy data, exhibited notable improvements in relevance and organization. Statistical analyses demonstrated some differences among note pools, with edited AI-generated notes consistently receiving the highest ratings. This highlights the importance of human oversight in enhancing AI output and tailoring it to specific clinical needs.

The research shows the potential of AI to augment clinical documentation processes, reduce clinician workload, and improve documentation quality. However, it also emphasizes the necessity of human-AI collaboration and robust training to mitigate limitations such as contextual inaccuracies and misclassifications. These findings provide a foundation for future research and practical recommendations for integrating AI into healthcare documentation workflows.

## Acknowledgements

I sincerely appreciate the invaluable support of my supervisors, Professor Maura R. Grossman and Professor Daniel G. Brown. Your generosity in funding my graduate studies, facilitating my Mitacs internship at the KidsAbility Centre for Child Development, and providing thoughtful guidance throughout my research journey has been instrumental in my academic and professional growth. From our first Zoom meeting on January 19, 2023, at 9:30 AM (EST), to completing this research, you have both demonstrated being the best supervisors anyone could ask for. You made time for me, responding at all hours (late nights, early mornings) and even during holidays and your leave. I want to say, “I am sorry” for not being able to call you by your first names as other students do during my time here. Maybe in the future? (*Just for laughs*). Thank you for your unwavering support and mentorship.

I would like to extend my thanks to Rachel DiMaio, the developer of KAUWbot, for her support and dedication to this project. Her innovative contributions to the development of KAUWbot were instrumental in advancing this research.

I am profoundly grateful to Brendan Wylie-Toal and the KidsAbility Centre for Child Development for their invaluable support and collaboration. Their provision of resources and dedication to advancing pediatric rehabilitation care played a vital role in making this research possible. I would like to express my sincere gratitude to Ilona Manda for her efforts in sorting the SOAP notes, managing the anonymization process, and coordinating the evaluation workflow, all of which were crucial to the smooth execution of this research. I also extend my appreciation to the four clinicians who generously contributed their time and expertise to evaluate the quality of the SOAP notes.

Thank you to Professor Helen Chen and Professor Jesse Hoey for reading my thesis and being part of the reason why my master’s program completion was successful.

I would like to thank Stella Mfoafo-M’Carthy of the University of Guelph and her husband, Professor Magnus Mfoafo-M’Carthy of Wilfrid Laurier University, for their support (emotional, financial, and always giving a listening ear and guidance when needed) in making sure I did not have a “lonely” graduate school experience.

To my former roommate (Oluwabamise “Bami” Afolabi) at Columbia Lake Village, you know why your name is here. *Na only you know wetin my eyes see for this school*. Thank you!

To my parents, family, and friends, pouring my heart out to you in appreciation would require more pages than my thesis. I appreciate and love you all, and you know it.

## Dedication

I dedicate this research to Ethan or Quinn, my first child (*yet to be born*).

# Table of Contents

|                                                                             |     |
|-----------------------------------------------------------------------------|-----|
| Author’s Declaration                                                        | ii  |
| Abstract                                                                    | iii |
| Acknowledgements                                                            | iv  |
| Dedication                                                                  | v   |
| List of Figures                                                             | ix  |
| List of Tables                                                              | xi  |
| 1 Introduction                                                              | 1   |
| 2 Background                                                                | 3   |
| 2.1 Clinical Documentation and SOAP Notes in Occupational Therapy . . . . . | 3   |
| 2.1.1 Pediatric Occupational Therapy . . . . .                              | 3   |
| 2.1.2 The Importance of Clinical Documentation in Occupational Therapy      | 4   |
| 2.1.3 SOAP Notes: Structure and Application . . . . .                       | 4   |
| 2.1.4 Limitations and Alternatives to SOAP Notes . . . . .                  | 5   |
| 2.1.5 Evolving Trends in Clinical Documentation . . . . .                   | 6   |
| 2.2 Artificial Intelligence in Clinical Documentation . . . . .             | 7   |
| 2.2.1 Ambient AI Scribes: Real-time Transcription and Summarization .       | 7   |
| 2.2.2 Large Language Models: Drafting and Generating Notes . . . . .        | 7   |
| 2.2.3 AI-Enhanced Electronic Health Record Systems: Integrated Solutions    | 8   |
| 2.2.4 Specialized Tools: Addressing Specific Needs . . . . .                | 8   |
| 2.2.5 Benefits of AI in Clinical Documentation . . . . .                    | 8   |

|          |                                                                                          |           |
|----------|------------------------------------------------------------------------------------------|-----------|
| 2.2.6    | Challenges and Limitations of AI in Clinical Documentation . . . . .                     | 9         |
| 2.3      | Overview of Microsoft Copilot as a General-Purpose LLM . . . . .                         | 10        |
| 2.4      | The Role of KAUWbot in Pediatric Rehabilitation Documentation . . . . .                  | 11        |
| 2.4.1    | About KidsAbility Centre for Child Development . . . . .                                 | 11        |
| 2.4.2    | The KAUWbot Concept . . . . .                                                            | 11        |
| 2.4.3    | KAUWbot Development: Data Acquisition, Processing, and Model Training . . . . .          | 11        |
| 2.4.4    | KAUWbot Pilot Study . . . . .                                                            | 12        |
| 2.5      | Evaluation of Clinical Documentation Quality . . . . .                                   | 13        |
| 2.5.1    | Methods for Assessing Clinical Documentation Quality . . . . .                           | 13        |
| 2.5.2    | Key Challenges in Achieving High-Quality Clinical Documentation . . . . .                | 14        |
| 2.5.3    | Strategies for Improving Clinical Documentation Quality . . . . .                        | 15        |
| 2.6      | Human and AI Collaboration in Clinical Documentation . . . . .                           | 16        |
| 2.6.1    | Human-AI Collaboration Models in Clinical Documentation . . . . .                        | 16        |
| 2.6.2    | Challenges and Limitations in Human-AI Collaboration in Clinical Documentation . . . . . | 17        |
| <b>3</b> | <b>Methodology</b>                                                                       | <b>18</b> |
| 3.1      | Study Population . . . . .                                                               | 18        |
| 3.2      | Data Collection . . . . .                                                                | 19        |
| 3.3      | SOAP Note Anonymization Process . . . . .                                                | 19        |
| 3.3.1    | Step 1: Assigning Letters to Each Pool . . . . .                                         | 19        |
| 3.3.2    | Step 2: Assigning Numbers to the Letters . . . . .                                       | 19        |
| 3.4      | From Scratch Notes to SOAP Notes . . . . .                                               | 20        |
| 3.5      | Human Editing Process of SOAP Notes . . . . .                                            | 21        |
| 3.6      | Rubric Criteria and Measurements . . . . .                                               | 21        |
| 3.7      | The Evaluation Process . . . . .                                                         | 22        |
| <b>4</b> | <b>Results and Analysis</b>                                                              | <b>24</b> |
| 4.1      | Mean Scores for the Four SOAP Note Pools . . . . .                                       | 24        |
| 4.2      | Quality Performance for the Four SOAP-Note Pools . . . . .                               | 27        |
| 4.2.1    | Descriptive Statistics of Quality Scores . . . . .                                       | 27        |
| 4.2.2    | ANOVA to Compare the Mean Quality Scores . . . . .                                       | 30        |
| 4.3      | Clarity Performance for the Four SOAP-Note Pools . . . . .                               | 32        |

|          |                                                                 |           |
|----------|-----------------------------------------------------------------|-----------|
| 4.4      | Completeness Performance for the Four SOAP-Note Pools . . . . . | 33        |
| 4.5      | Conciseness Performance for the Four SOAP Note Pools . . . . .  | 37        |
| 4.6      | Relevance Performance for the Four SOAP-Note Pools . . . . .    | 38        |
| 4.7      | Organization Performance for the Four SOAP-Note Pools . . . . . | 41        |
| <b>5</b> | <b>Discussion</b>                                               | <b>43</b> |
| 5.1      | Quality of AI-Generated SOAP Notes . . . . .                    | 43        |
| 5.2      | Human-AI Collaboration . . . . .                                | 44        |
| 5.3      | Implications for Clinical Practice . . . . .                    | 44        |
| 5.4      | Limitations and Recommendations for Future Research . . . . .   | 44        |
| <b>6</b> | <b>Conclusion and Summary</b>                                   | <b>46</b> |
| 6.1      | Summary of Key Findings . . . . .                               | 46        |
|          | <b>References</b>                                               | <b>47</b> |

# List of Figures

|      |                                                                                                       |    |
|------|-------------------------------------------------------------------------------------------------------|----|
| 3.1  | Sample Prompt for Copilot, Sample Scratch Note and Generated Copilot SOAP Note [5]. . . . .           | 20 |
| 3.2  | SOAP Note Evaluation Instrument. . . . .                                                              | 23 |
| 4.1  | Histogram of Clinician Scoring of Human-Authored SOAP Notes. . . . .                                  | 25 |
| 4.2  | Histogram of Clinician Scoring of Copilot-Generated SOAP Notes Edited by OTs. . . . .                 | 26 |
| 4.3  | Histogram of Clinician Scoring of KAUWbot-Generated SOAP Notes Unedited. . . . .                      | 26 |
| 4.4  | Histogram of Clinician Scoring of KAUWbot-Generated SOAP Notes Edited by OTs. . . . .                 | 28 |
| 4.5  | Box Plot of SOAP Note Quality Scores by Evaluators . . . . .                                          | 29 |
| 4.6  | Violin Plot of SOAP Note Quality Scores by Evaluators . . . . .                                       | 29 |
| 4.7  | Histogram of Mean Quality Scores for SOAP-Note Pools [5]. . . . .                                     | 30 |
| 4.8  | Histogram of Mean, Median, and Standard Deviation of Quality Scores for SOAP-Note Pools. . . . .      | 31 |
| 4.9  | Histogram of Mean Clarity Scores for SOAP-Note Pools. . . . .                                         | 34 |
| 4.10 | Histogram of Mean, Median, and Standard Deviation of Clarity Scores for SOAP-Note Pools. . . . .      | 34 |
| 4.11 | Histogram of Mean Completeness Scores for SOAP-Note Pools. . . . .                                    | 36 |
| 4.12 | Histogram of Mean, Median, and Standard Deviation of Completeness Scores for SOAP-Note Pools. . . . . | 36 |
| 4.13 | Histogram of Mean Conciseness Scores for SOAP-Note Pools. . . . .                                     | 38 |
| 4.14 | Histogram of Mean, Median, and Standard Deviation of Conciseness Scores for SOAP-Note Pools. . . . .  | 38 |
| 4.15 | Histogram of Mean Relevance Scores for SOAP-Note Pools. . . . .                                       | 40 |
| 4.16 | Histogram of Mean, Median, and Standard Deviation of Relevance Scores for SOAP-Note Pools. . . . .    | 40 |
| 4.17 | Histogram of Mean Relevance Scores for SOAP-Note Pools. . . . .                                       | 42 |

|                                                                                                               |    |
|---------------------------------------------------------------------------------------------------------------|----|
| 4.18 Histogram of Mean, Median, and Standard Deviation of Organization Scores<br>for SOAP-Note Pools. . . . . | 42 |
|---------------------------------------------------------------------------------------------------------------|----|

# List of Tables

|      |                                                                                                          |    |
|------|----------------------------------------------------------------------------------------------------------|----|
| 3.1  | The Four SOAP Note Pools and Their Description . . . . .                                                 | 18 |
| 4.1  | Human-Authored SOAP Notes (K) . . . . .                                                                  | 24 |
| 4.2  | Copilot SOAP Notes - OT edited (E) . . . . .                                                             | 25 |
| 4.3  | KAUWbot SOAP Notes - unedited (U) . . . . .                                                              | 27 |
| 4.4  | KAUWbot SOAP Notes - OT edited (T) . . . . .                                                             | 27 |
| 4.5  | Summary of Mean Values for Evaluators' (Clinician) Quality Scores for the SOAP-Note Pools . . . . .      | 27 |
| 4.6  | Summary of Data for Evaluators' (Clinician) Quality Scores for the SOAP-Note Pools . . . . .             | 28 |
| 4.7  | Summary of Analysis of Variance (ANOVA) with $\alpha = 0.05$ . . . . .                                   | 30 |
| 4.8  | Summary of Mean Values for Evaluators' (Clinician) Clarity Scores for the SOAP-Note Pools . . . . .      | 33 |
| 4.9  | Summary of Data for Evaluators' (Clinician) Clarity Scores for the SOAP-Note Pools . . . . .             | 33 |
| 4.10 | Summary of Mean Values for Evaluators' (Clinician) Completeness Scores for the SOAP-Note Pools . . . . . | 35 |
| 4.11 | Summary of Data for Evaluators' (Clinician) Completeness Scores for the SOAP-Note Pools . . . . .        | 35 |
| 4.12 | Summary of Mean Values for Evaluators' (Clinician) Conciseness Scores for the SOAP-Note Pools . . . . .  | 37 |
| 4.13 | Summary of Data for Evaluators' (Clinician) Conciseness Scores for the SOAP-Note Pools . . . . .         | 37 |
| 4.14 | Summary of Mean Values for Evaluators' (Clinician) Relevance Scores for the SOAP-Note Pools . . . . .    | 39 |
| 4.15 | Summary of Data for Evaluators' (Clinician) Relevance Scores for the SOAP-Note Pools . . . . .           | 39 |
| 4.16 | Summary of Mean Values for Evaluators' (Clinician) Organization Scores for the SOAP-Note Pools . . . . . | 41 |

|                                                                                                           |    |
|-----------------------------------------------------------------------------------------------------------|----|
| 4.17 Summary of Data for Evaluators' (Clinician) Organization Scores for the<br>SOAP-Note Pools . . . . . | 41 |
|-----------------------------------------------------------------------------------------------------------|----|

# Chapter 1

## Introduction

Artificial intelligence (AI), when integrated into healthcare, can transform various aspects of clinical practice, including diagnostics, treatment planning, and administrative tasks [3]. One emerging application is the automation of clinical documentation through AI-generated SOAP (Subjective, Objective, Assessment, and Plan) notes. These structured notes are essential for tracking patient progress, facilitating communication among healthcare providers, and supporting legal and billing requirements [7, 19]. In pediatric rehabilitation settings, documentation demands are particularly high due to the complexity and individualization of treatment plans [81, 89]. AI-powered tools like Microsoft’s Copilot and KAUWbot (a custom AI tool developed through the collaboration between the KidsAbility Centre for Child Development and the University of Waterloo to generate pediatric rehabilitation SOAP notes) can simplify these processes, reducing clinician workload and allowing more time for direct patient care [39]. However, concerns remain about AI-generated notes’ quality, reliability, and clinical utility [5]. This research evaluates these aspects, focusing on pediatric rehabilitation SOAP notes, to determine whether AI tools can meet or exceed current documentation standards.

Despite AI’s potential to revolutionize clinical documentation, the quality of AI-generated SOAP notes in pediatric rehabilitation remains uncertain [5]. Rigorous evaluation is essential to ensure these tools meet established quality standards. Without such evaluation, adopting these technologies may introduce risks to clinical accuracy, patient safety, and continuity of care [2].

This research addresses two key questions: First, how does the quality of AI-generated pediatric rehabilitation SOAP notes compare to human-authored notes in terms of *clarity*, *completeness*, *conciseness*, *relevance*, and *organization*? Based on this research’s five custom quality criteria, this question evaluates whether AI-generated notes meet or exceed established clinical documentation standards. The second question asks to what extent human editing is necessary and how much human editing is required for SOAP notes generated by different AI tools. This question examines the extent of human intervention required for AI-generated notes to reach acceptable quality standards, comparing the performance of Copilot and KAUWbot.

The research objectives are to compare the quality of human-authored SOAP notes with

AI-generated SOAP notes, evaluate the necessity and amount of human editing required to achieve acceptable quality SOAP notes, and provide insights into the potential benefits and limitations of using AI in clinical documentation. This research offers valuable insights into the practical application of AI in clinical documentation. By assessing the quality and required human editing of AI-generated SOAP notes, the research provides evidence-based recommendations for healthcare institutions considering adopting such technologies. Understanding how different AI models perform could guide investment decisions and inform best practices for integrating AI into clinical workflows. The findings could also contribute to broader discussions on AI ethics, data quality, and clinician-AI collaboration, promoting more effective and safe healthcare documentation systems.

This thesis is structured into six chapters. Chapter 1 introduces the study by presenting the context, problem statement, research questions, objectives, and significance of the research. Chapter 2 delves into a detailed review of the relevant literature, focusing on AI's role in clinical documentation, particularly on AI-generated SOAP notes and their implications in pediatric rehabilitation. Chapter 3 outlines the research methodology, describing the processes involved in data collection, anonymization, and evaluation. Chapter 4 presents the results and analysis, showcasing key findings from the rubric-based evaluations and statistical tests, including ANOVA analysis. Chapter 5 discusses the implications of these findings, contextualizing them within the broader literature and highlighting the strengths, limitations, and potential applications of AI in pediatric clinical documentation; proposing practice recommendations, and identifying areas for future research. Chapter 6 concludes the thesis by summarizing the key contributions.

# Chapter 2

## Background

This chapter provides a review of the literature relevant to clinical documentation in occupational therapy (OT) and the role of AI in enhancing documentation quality. It begins with an overview of clinical documentation practices, specifically focusing on the use of SOAP notes in OT. Next, it examines the integration of AI in clinical documentation, including an overview of Microsoft Copilot as a general-purpose large language model (LLM) and the role of KAUWbot in supporting pediatric rehabilitation documentation. Additionally, this chapter explores existing methods for evaluating clinical documentation quality and discusses the dynamics of human-AI collaboration in clinical documentation.

### 2.1 Clinical Documentation and SOAP Notes in Occupational Therapy

Clinical documentation is a cornerstone of OT practice, serving multiple purposes including patient care, reimbursement, research, and professional development [19]. A significant portion of this documentation involves the creation of SOAP notes, a structured method for recording client information. This section describes pediatric occupational therapy and examines the role of clinical documentation, particularly SOAP notes, within OT, exploring their structure, applications, limitations, and evolving alternatives.

#### 2.1.1 Pediatric Occupational Therapy

Pediatric occupational therapy is a specialized area of practice focused on enhancing the participation and well-being of children and youth with diverse developmental needs [88, 1, 100]. It is a client-centered approach that emphasizes the child's unique abilities and challenges within their natural environment [52, 32]. Unlike some misconceptions that limit intervention to a single approach like Ayres Sensory Integration [88], pediatric OT adopts a multifaceted strategy, incorporating various evidence-based methods to address the whole child across varied contexts [88, 1]. This comprehensive understanding considers

the child’s physical, cognitive, social-emotional, and sensory-motor skills within their family and community settings [22, 76, 87]. The goal is to improve the child’s ability to engage in meaningful occupations – activities that are essential to their daily lives and personal fulfillment [100, 63, 44].

### 2.1.2 The Importance of Clinical Documentation in Occupational Therapy

Effective clinical documentation is crucial for several reasons. Firstly, it comprehensively records the patient’s history, assessment findings, intervention plans, and progress [7]. This detailed record is essential for ongoing care, facilitating communication between health-care professionals and ensuring continuity of treatment. Secondly, comprehensive assessments, detailed in the documentation, allow occupational therapists (OTs) to identify client strengths and weaknesses, establish goals, and tailor interventions to meet individual needs. The ongoing documentation of client progress allows for modifications to the treatment plan, ensuring its ongoing relevance and effectiveness [28, 105]. Furthermore, clinical documentation is a valuable resource for research, enabling the analysis of treatment outcomes and the development of evidence-based practices [46]. Finally, comprehensive and accurate documentation provides crucial legal protection for OT practitioners [4]. It serves as a defense against potential malpractice claims by providing a clear record of the assessment, treatment, and progress of each client [4].

The importance of documentation is emphasized by the fact that it is learned as a skill, requiring both instruction and practice [19]. Several resources exist to aid in the development of these documentation skills, ranging from textbooks offering templates and practice worksheets [19, 49] to instructor manuals with additional assignments and supplementary videos [49].

### 2.1.3 SOAP Notes: Structure and Application

The SOAP note format, an acronym for Subjective, Objective, Assessment, and Plan, is a widely adopted method for documenting client interactions in OT [49, 7].

#### Subjective (S):

This section captures the client’s self-reported information, including their feelings, perceptions, and concerns regarding their occupational performance [49]. It is crucial to document the client’s perspective accurately, reflecting their lived experience and understanding of their challenges [31]. The subjective section is vital for person-centered care, ensuring the client’s voice is central to the treatment process [31]. However, difficulties in accurately capturing a child’s self-reported symptoms due to age-related communication limitations [24], the need to rely heavily on caregiver input [93], and interpreting non-verbal cues [21] are some challenges for this section in pediatric rehabilitation.

**Objective (O):**

The objective section presents measurable and observable data obtained during the OT session [49]. This includes quantitative data such as range of motion, strength measurements, or time taken to complete a task, as well as qualitative observations of the client's performance and behaviours [7]. The objective section must be detailed and specific, avoiding vague or subjective descriptions. The use of standardized assessment tools can enhance the reliability and validity of objective data [68].

**Assessment (A):**

The assessment section integrates the subjective and objective information to form a clinical interpretation of the client's status [49]. It analyzes the client's strengths and weaknesses, identifies contributing factors to their difficulties, and determines the overall impact of their occupational challenges on their daily life [49]. The assessment section should be concise yet thorough, clearly articulating the therapist's professional judgment and rationale for treatment decisions [49].

**Plan (P):**

The plan section outlines the proposed intervention strategies for the next session or treatment period [49]. It should specify the goals, techniques, and expected outcomes, aligning with the assessment and the client's individual needs [7]. The plan should be flexible and adaptable, allowing for modifications based on the client's progress and response to intervention [7].

### 2.1.4 Limitations and Alternatives to SOAP Notes

While SOAP notes have been a mainstay of OT documentation, they are not without limitations. The primary criticism centres on the potential for brevity and lack of context, which can hinder the comprehensive understanding of the client's situation [15]. The focus on discrete sections can lead to fragmented information and a loss of the holistic perspective inherent in OT practice. The subjective section is important but does not always fully represent the client's experience, which can lead to misinterpretations. Some researchers have suggested that the SOAP note format's emphasis on physical skills in self-care tasks can overshadow the unique philosophies and approaches central to OT [15].

In response to these limitations, alternative documentation methods are being explored and implemented. The shift towards person-centered care has highlighted the need for more narrative and holistic approaches [31]. These approaches focus on the client's story and lived experience, offering a deeper understanding of their occupational performance. Furthermore, the incorporation of the International Classification of Functioning, Disability, and Health (ICF) framework into documentation has become increasingly prevalent, offering a standardized and comprehensive language for describing functional limitations and

their impact on participation [15]. The ICF provides a common language for professionals across various disciplines, enhancing interprofessional communication and collaboration [7]. The use of consult notes, a more succinct documentation style, is being explored as a potential alternative, particularly in settings like pharmacy where efficiency is important [18].

### 2.1.5 Evolving Trends in Clinical Documentation

Several trends are shaping the future of clinical documentation in OT. The increasing use of electronic health records (EHRs) is transforming how information is stored, accessed, and shared [97, 7]. EHRs offer the potential for improved efficiency, reduced errors, and enhanced communication among healthcare professionals. However, the implementation of EHRs also presents challenges, including the potential for “note bloat” and the need for effective strategies to manage and analyze the large volumes of data generated [97, 45]. The development of novel note templates, designed to simplify information and reduce unnecessary detail, is one approach to address this issue [45]. Nonetheless, over-reliance on templates may lead to the exclusion of pertinent patient details not covered by the template’s structure, potentially compromising patient care [109]. The convenience of copying and pasting template text can result in outdated or incorrect information being carried forward, especially if records are not carefully updated during each patient encounter [33].

Another important trend is the growing emphasis on data-driven decision-making and outcome measurement [64, 68]. The use of standardized assessment tools and outcome measures allows for the systematic evaluation of treatment effectiveness and the identification of areas for improvement. This data-driven approach is crucial for ensuring that OT interventions are evidence-based and effective in achieving desired outcomes [46].

The integration of technology, such as virtual reality (VR) and telehealth, is also impacting clinical documentation. VR-based interventions, for instance, require specific documentation of the client’s experience and progress within the virtual environment [85]. Similarly, telehealth sessions necessitate documentation of remote interactions, technical challenges, and adaptations to the therapeutic process. The development of standardized methods for documenting these innovative approaches is an ongoing area of development [85].

The focus on person-centered care continues to influence documentation practices, emphasizing the importance of recording the client’s voice and perspective [31]. This requires therapists to be skilled in eliciting and documenting subjective experiences, ensuring that the client’s goals and values are central to the treatment process. The use of collaborative documentation methods, involving the client and family in the process, can further enhance person-centered care [31].

Finally, the legal and ethical considerations surrounding clinical documentation remain crucial [7]. Therapists must adhere to strict confidentiality guidelines, ensuring the protection of sensitive client information [7]. Accurate and thorough documentation is also essential for avoiding legal action and demonstrating compliance with regulatory standards.

This requires a strong understanding of relevant laws and ethical principles, including those related to record retention and data security [7].

## 2.2 Artificial Intelligence in Clinical Documentation

The increasing volume and complexity of clinical documentation pose significant challenges to healthcare professionals. Clinician burnout, stemming from excessive administrative burdens [81, 89], is a major concern. Maintaining accurate and comprehensive patient records is key for effective communication, continuity of care, and overall patient safety [14, 10]. AI, and particularly LLMs and natural language processing (NLP) techniques, offers potential solutions to simplify clinical documentation, improve its quality, and alleviate the burden on healthcare providers [27, 75]. This section examines the current state of AI applications in clinical documentation, exploring both the promises and challenges of this rapidly evolving field.

### 2.2.1 Ambient AI Scribes: Real-time Transcription and Summarization

Ambient AI scribes utilize automatic speech recognition (ASR) and NLP to transcribe and summarize doctor-patient conversations in real-time [70]. These systems aim to eliminate manual note-taking, freeing clinicians to focus on patient interaction. One example is Sporo Health’s AI scribe, a multi-agent system leveraging fine-tuned medical LLMs [70], which has demonstrated superior performance compared to GPT-4o Mini in terms of clinical content recall, precision, and F1 scores [70]. However, challenges remain, including ensuring accuracy in handling complex medical language and maintaining patient data privacy [70]. The impact of non-lexical conversational sounds (NLCS), such as ”mm-hm” and ”uh-huh,” on ASR performance also requires further investigation [110]. Furthermore, the successful implementation of ambient AI scribes is contingent upon factors like flexible integration into existing workflows, support for clinician sensemaking, and the ability to experiment with the AI [17].

### 2.2.2 Large Language Models: Drafting and Generating Notes

LLMs like ChatGPT and GPT-4 are increasingly being used to generate drafts of clinical notes, discharge summaries, and other documentation [82, 69, 14, 95]. These models can process large amounts of text data and generate coherent, grammatically correct text within seconds [82]. A commentary from a clinical and research perspective highlights the benefits and pitfalls of implementing AI-driven models for the creation and management of healthcare progress notes [82]. Researchers have successfully used prompt engineering techniques to generate notes in various contexts, such as chronic pain assessments and psychiatric evaluations [82]. However, concerns remain regarding the accuracy and reliability

of LLM-generated content, which is heavily dependent on the quality of input data and the effectiveness of prompting [82].

The potential for inherent biases in the models trained on internet data needs to be carefully addressed [82, 99]. A study comparing Sporo AI Scribe and GPT-4o Mini found that Sporo AI consistently outperformed GPT-4o Mini across several performance metrics [70], highlighting the importance of fine-tuning and specialized medical LLMs. The use of LLMs for drafting responses to patient messages is also being explored [99], with some studies suggesting potential improvements in efficiency and patient education [27]. However, concerns persist about the risk of harm from unedited AI-generated responses [27].

### **2.2.3 AI-Enhanced Electronic Health Record Systems: Integrated Solutions**

Several AI-powered tools are being integrated directly into EHR systems to enhance documentation workflows [90, 74]. These tools may automate data entry, suggest relevant codes, or provide other forms of clinical decision support [47]. A non-randomized clinical trial investigated the EHR experiences of clinicians before and after the implementation of an AI-powered clinical documentation tool [74]. However, the integration of AI into EHRs presents challenges related to interoperability, data security, and the need for robust validation and control processes [82]. The successful use of AI within EHRs requires AI-capable organizations with the necessary internal and external systems [84].

### **2.2.4 Specialized Tools: Addressing Specific Needs**

Beyond general-purpose documentation tools, AI is also being applied to address specific needs in various medical subspecialties. For instance, in hepatology, AI algorithms are being implemented for liver imaging interpretation, histopathology analysis, and non-invasive test interpretation [92]. In radiology, AI-powered tools are being developed for the detection of polyps during colonoscopy [38] and the automated assessment of body composition from CT images [16]. In emergency medicine, AI is being explored for automated triage, diagnostic support, and predictive analytics [71, 29, 96]. The application of AI in dental radiology is also being investigated for generating reports from panoramic radiographs [101].

### **2.2.5 Benefits of AI in Clinical Documentation**

AI is transforming clinical documentation by making the process more efficient and freeing up valuable time for healthcare providers. Tasks like transcription, data entry, and report generation, which are often tedious and time-consuming, can now be automated [14, 81]. This allows clinicians to focus more on their patients, increasing productivity and reducing the likelihood of burnout [14]. Beyond saving time, AI also improves the accuracy and

completeness of documentation by catching inconsistencies, flagging issues, and ensuring nothing important gets overlooked [89, 70].

Several experiments have demonstrated the impact of AI in improving communication, documentation efficiency, and cost savings in healthcare settings. AI-generated reports and summaries enhance coordination among healthcare providers by ensuring all team members can access clear, standardized information [10, 56]. The findings suggest that AI-driven documentation reduces ambiguity and improves collaboration, ultimately leading to better patient outcomes [10, 56]. Similarly, AI’s role in simplifying teamwork within clinical environments shows that structured and consistent documentation facilitates smoother decision-making processes [14, 27].

In addition to communication improvements, AI’s potential for cost reduction in healthcare documentation highlights that AI minimizes the need for manual data entry, allowing clinicians to dedicate more time to patient care [81, 11]. By automating routine documentation tasks, AI improves efficiency and reduces administrative burdens and associated costs for healthcare organizations [81, 11].

### **2.2.6 Challenges and Limitations of AI in Clinical Documentation**

AI-powered clinical documentation tools present a range of challenges that must be addressed to ensure their effective and ethical use in healthcare. One key concern is accuracy and reliability, as AI systems, particularly LLMs, can sometimes produce inaccurate or incomplete information [10]. Another challenge involves bias and fairness. Since AI systems are trained on datasets that may contain biases, there is a risk of perpetuating unfair or discriminatory outcomes, particularly for marginalized patient populations [8, 80]. The use of race-based algorithms in medicine, for example, has been shown to perpetuate health inequities [25], highlighting the need for a race-conscious approach that prioritizes addressing systemic racism rather than relying on race as a proxy for biological differences [25].

Privacy and security are also major concerns. Clinical data is highly sensitive, and any breaches could result in significant legal, public relations, and ethical ramifications [81]. The lack of explainability and transparency in some AI systems, particularly those based on deep learning, raises trust issues. These systems often operate as “black boxes,” making it difficult to understand how decisions are made or to identify errors [8, 20]. Integrating AI tools into existing EHR systems also poses technical challenges, as differences in architectures and data formats can complicate seamless integration [30]. A notable incident occurred in 2017 when DeepMind, a Google-owned company, was found to have not complied with UK data protection laws during its partnership with the UK’s National Health Service (NHS) [61]. The Information Commissioner’s Office determined that patients were not adequately informed about how their data was being utilized, leading to concerns over unauthorized access and potential misuse of sensitive health information [61]. In another case, Babylon Health, a digital health service provider, experienced a data breach in June 2020 where patients were inadvertently given access to video recordings of other patients’

consultations [59]. This breach highlighted vulnerabilities in AI-driven telehealth platforms and raised questions about the adequacy of their security measures.

Clinicians may be reluctant to adopt AI-driven tools if they have concerns about their accuracy, reliability, or usability. Providing adequate training and support is needed for promoting user acceptance [94]. The legal and ethical implications of using AI in clinical documentation, such as liability for errors, data ownership, and informed consent, remain unresolved. Clear guidelines and regulations are needed to address these issues [34, 72].

## 2.3 Overview of Microsoft Copilot as a General-Purpose LLM

Microsoft Copilot’s primary function was to provide real-time code suggestions and completions within integrated development environments (IDEs) [106]. Trained on a massive dataset of publicly available code, it could generate code snippets, entire functions, and even complete programs based on natural language descriptions or partial code input [37]. This functionality significantly boosted developer productivity by automating repetitive coding tasks and suggesting optimal solutions [43].

Beyond code generation, Copilot’s LLM capabilities have been extended to other tasks. For instance, it can now generate documentation, translate code between programming languages, and assist with debugging [106]. This broader functionality illustrates its potential as a more general-purpose tool within the software development lifecycle [66]. In healthcare, for example, Copilot has shown promise in answering frequently asked questions related to breast imaging, demonstrating potential improvements in health literacy and patient engagement [108]. However, studies also note variations in accuracy and readability across different LLMs, including Copilot, highlighting the need for ongoing evaluation and refinement [108]. Similarly, in the biomedical field, the BioSignal Copilot uses LLMs to draft reports for biomedical signals, addressing the gap in applying LLMs to time-series data [73].

In other sectors, research explores the use of Copilot for tasks such as generating PLC/DCS control logic [67], assisting with model-driven tasks using UML [23], and even enhancing the design of future vehicles through the integration of Retrieval-Augmented Generation (RAG) [115]. These diverse applications show the versatility of Copilot’s LLM foundation, making it adaptable to various contexts and problem-solving scenarios. The integration of LLMs with RAG, as explored in the automotive industry, promises improved accuracy and contextual awareness in automated design and software development [115].

The educational sector has also witnessed the integration of Copilot, with studies examining its impact on student learning outcomes [55, 57]. While Copilot can be a valuable tool for knowledge discovery and assistance [57], concerns exist regarding over-reliance and potential negative impacts on independent problem-solving skills [55]. The findings highlight the importance of responsible integration and the need for educators to guide students in the ethical and effective use of such tools [55].

## 2.4 The Role of KAUWbot in Pediatric Rehabilitation Documentation

This section provides an overview of the development and evaluation of KAUWbot. It begins with a description of KidsAbility Centre for Child Development, the clinical setting in which KAUWbot was conceptualized. The subsequent subsections outline the KAUWbot concept, the processes involved in data acquisition, processing, and model training, and findings from a pilot study assessing its effectiveness. This section serves as a summary of the master’s thesis, *Creation of a Custom Language Model for Pediatric Occupational Therapy Documentation* [39], which details the research and development efforts behind KAUWbot.

### 2.4.1 About KidsAbility Centre for Child Development

KidsAbility Centre for Child Development is a pediatric rehabilitation organization serving over 17,000 children and youth annually in the Waterloo and Wellington regions of Ontario, Canada [62]. Established in 1957 by local Rotarians, it has expanded to multiple locations, including Waterloo, Kitchener, Guelph, Fergus, and Mount Forest, to enhance accessibility for families. The Centre offers a range of services, such as occupational therapy, speech-language therapy, physiotherapy, social work, and autism and behavior supports, all aimed at helping clients achieve milestones in communication, social, physical, and behavioral development [62]. Emphasizing early, family-centered interventions, KidsAbility supports children from birth to age 21, guiding them toward realizing their full potential [62].

### 2.4.2 The KAUWbot Concept

Recent advances in LLMs have shown potential for automating various tasks in healthcare, including clinical documentation. KidsAbility’s research and innovation team recognized the potential of LLMs to address their documentation challenges [39]. Initial explorations with ChatGPT yielded promising results [39], but concerns regarding data privacy and security, given the proprietary nature of ChatGPT and the potential for exposure of personal health information (PHI), led to the decision to develop a custom model. This custom model, KAUWbot, would be trained on KidsAbility’s data, ensuring both privacy and better alignment with the specific terminology and style preferences of its clinicians [39].

### 2.4.3 KAUWbot Development: Data Acquisition, Processing, and Model Training

The development of KAUWbot involved several key stages: data acquisition, data processing, and model training. KidsAbility provided two datasets of historical progress notes in

SOAP format. These notes contained incidental PHI, primarily first names and ages, which needed to be removed to ensure patient privacy. Amazon Web Services Comprehend was utilized for automated redaction, although manual review was necessary to ensure complete removal of PHI. The processed datasets, containing thousands of redacted SOAP notes, formed the basis for KAUWbot’s training [39].

Because the initial datasets lacked corresponding scratch notes (the point-form notes taken during therapy sessions), a synthetic dataset of scratch notes was generated using the Llama 2 70B Chat model. A human-in-the-loop evaluation process, involving KidsAbility clinicians using a rubric to assess the quality of the generated scratch notes, was implemented to refine the generation procedure. This iterative process ensured that the synthetic scratch notes were realistic and representative of actual clinician notes. The generated scratch notes were then paired with the corresponding historical SOAP notes to create a training dataset for KAUWbot [39].

Several training techniques were explored, including domain-adaptive pre-training (DAPT) and low-rank adaptation (LoRA) fine-tuning. The Meta Llama models, specifically the smaller models (*7-8 billion parameters*), were chosen for their accessibility and proficiency in NLP tasks. Different Llama models underwent various training paradigms, and the best-performing model, a Llama 3 8B Instruct model fine-tuned on the progress notes and generated scratch notes, was selected for the pilot study. The choice of smaller, open-source models also addressed concerns about computational resources and the environmental impact of training large proprietary models. The use of open-source models allowed for hosting on KidsAbility’s private servers, mitigating data privacy risks [39].

#### 2.4.4 KAUWbot Pilot Study

A pilot study was conducted at KidsAbility to evaluate KAUWbot’s effectiveness in reducing documentation time. Ten occupational therapy clinicians participated, using both Microsoft Copilot and KAUWbot to write progress notes over three-week periods. The study included qualitative and quantitative data collection methods. Qualitative data was gathered through clinician feedback, which provided valuable insights into usability and areas for improvement. Quantitative data included ROUGE scores to compare the similarity between edited and unedited notes generated by KAUWbot, and timing data to assess the impact on documentation time [39].

The time savings achieved with KAUWbot varied significantly among clinicians. This variability was largely attributed to differences in how clinicians used the model and the detail level of their scratch notes. Clinicians who provided sparser scratch notes experienced greater efficiency and ease in generating final SOAP notes. The importance of clinician training on effective usage of the model was highlighted, with coaching leading to significant time reductions for some clinicians. This training focused on providing concise scratch notes and effectively utilizing the model’s capabilities to generate a first draft that requires minimal editing [39].

Following the development and implementation of KAUWbot, clinicians at KidsAbility were able to generate SOAP notes in less time than those who documented without the AI

system. While this represents a significant advancement in workflow efficiency, it remains unclear whether the SOAP notes produced by KAUWbot meet acceptable clinical quality standards [5].

## 2.5 Evaluation of Clinical Documentation Quality

High-quality clinical documentation is paramount to effective healthcare delivery. It facilitates clear communication among healthcare professionals, ensures patient safety, supports accurate coding and billing, and provides crucial data for research and quality improvement initiatives [112, 9, 51]. Achieving and maintaining high-quality documentation presents significant challenges across various healthcare settings. This section systematically examines the current state of clinical documentation quality, exploring diverse methodologies for assessment, identifying key challenges, and evaluating strategies for improvement.

### 2.5.1 Methods for Assessing Clinical Documentation Quality

The assessment of clinical documentation quality employs a variety of approaches, ranging from simple checklist-based audits to sophisticated NLP techniques [13, 35, 113]. Studies often focus on specific aspects of documentation, such as completeness, accuracy, readability, and adherence to established guidelines [50, 48, 79].

#### Checklist-Based Audits

Many studies use checklists to evaluate the presence or absence of specific data elements within medical records [113, 79]. These checklists can be tailored to specific specialties or clinical contexts, focusing on key aspects of documentation relevant to patient care, coding, or legal requirements [79]. While straightforward to implement, checklist-based audits may not fully capture the complexities of clinical language or the overall quality of narrative descriptions [40].

#### Quantitative Analysis

Quantitative methods, such as calculating the percentage of completed items on a checklist or measuring the length of documentation, provide objective measures of documentation quality [35, 113, 79]. These approaches are useful for tracking progress over time and comparing documentation quality across different units or specialties [58, 42]. However, quantitative measures alone may not fully capture the qualitative aspects of documentation, such as clarity and organization [40].

## Qualitative Analysis

Qualitative analysis involves examining the content of medical records to assess clarity, organization, and the presence of essential clinical information [40, 13]. This approach can provide valuable insights into the reasons for poor documentation quality and identify areas for improvement that quantitative methods might miss [40]. However, qualitative analysis is more time-consuming and subjective than quantitative methods [13].

## Natural Language Processing Techniques

NLP techniques are increasingly being used to analyze unstructured clinical text data, providing more sophisticated methods for assessing documentation quality [13, 14]. These techniques can identify patterns in language use, detect inconsistencies, and assess the completeness and accuracy of clinical information [14]. However, NLP approaches can be complex to implement and require significant computational resources [13]. NLP has been employed to identify and correct errors in dictated medical documents, thereby improving their accuracy and reliability [6].

### 2.5.2 Key Challenges in Achieving High-Quality Clinical Documentation

Achieving high-quality clinical documentation is essential for effective patient care, yet several challenges hinder this goal. This section provides an overview of some of the challenges, such as time constraints, the complexity of medical language, inadequate training, inefficient system design, and a lack of standardization that contribute to inconsistencies and errors in documentation.

#### Time Constraints

Occupational therapists often face significant time pressures, limiting their ability to dedicate sufficient time to thorough documentation [98]. This is particularly true in busy clinical settings where the focus may shift towards immediate patient needs rather than detailed record-keeping [98].

#### Complexity of Medical Language

The complexity of medical terminology and the need for precise language can make documentation challenging [40, 14]. Inconsistent use of terminology can lead to ambiguity and errors, compromising the quality of clinical information [40].

## **Inadequate Training**

Many healthcare professionals receive insufficient training in effective clinical documentation techniques [91]. This lack of training can result in incomplete, inaccurate, or poorly organized records [91].

## **System Design and Workflows**

The design of EHR systems and clinical workflows can significantly impact documentation quality [54]. Poorly designed systems or inefficient workflows can hinder efficient and effective documentation [54].

## **Lack of Standardization**

The absence of standardized documentation practices or templates can lead to inconsistencies in the quality and completeness of clinical information [42, 79]. This lack of standardization can make it difficult to compare data across different providers or healthcare settings [42].

### **2.5.3 Strategies for Improving Clinical Documentation Quality**

Strategies such as educational interventions, template-based documentation, and feedback mechanisms as discussed in this section help standardize documentation practices and reduce errors.

#### **Educational Interventions**

Educational interventions, such as workshops, training sessions, or online modules, can significantly improve healthcare professionals' knowledge and skills in documentation [91, 58]. These interventions often focus on specific aspects of documentation, such as the use of standardized terminology, the importance of accurate recording of clinical findings, and the implementation of effective documentation techniques [91].

#### **Template-Based Documentation**

The use of standardized templates or forms can enhance documentation consistency and completeness [42, 79]. These templates often include prompts for key data elements, ensuring that essential clinical information is consistently captured [42].

## Feedback Mechanisms

Providing regular feedback to healthcare professionals on their documentation quality can promote continuous improvement [58]. This feedback can be provided through chart reviews, audits, or other mechanisms [58].

## Clinical Documentation Improvement (CDI) Programs

CDI programs employ dedicated specialists to work with clinicians to improve documentation quality [104, 51]. These programs typically involve proactive querying of clinicians to clarify ambiguous or incomplete information [104].

## Technology-Based Solutions

Technology-based solutions, such as speech recognition software or AI-powered tools, can automate aspects of documentation, reducing the time and effort required for record-keeping [14, 60]. These tools can also help to improve the accuracy and consistency of documentation [14].

## 2.6 Human and AI Collaboration in Clinical Documentation

Human-AI collaboration models aim to use AI’s capabilities alongside human expertise to enhance the quality and consistency of clinical documentation. Despite these advancements, several challenges and limitations remain. This section explores different models of human-AI collaboration in clinical documentation and examines the key obstacles that impact their effectiveness in real-world healthcare settings.

### 2.6.1 Human-AI Collaboration Models in Clinical Documentation

Several distinct models of human-AI collaboration in documentation have emerged, each with its strengths and limitations. One common model is the “helper” model, where AI acts as a tool to assist human documenters. This is exemplified by the use of Computer Assisted Coding (CAC) systems by Clinical Documentation Integrity Specialists (CDIS) [17]. In this scenario, the AI system automates tasks such as code suggestion, freeing up the human to focus on higher-level tasks requiring judgment and critical thinking. The CDIS, acting as “Batman,” uses the AI’s “Robin” assistance, recognizing the AI’s inherent limitations while still appreciating its value [17].

Another model involves a more integrated and symbiotic relationship, where humans and AI work together to bring out efficient output. This approach is explored in the

context of primary care consultations [65], where an AI documentation assistant could record and structure information gathered during the consultation. However, this model necessitates careful design to maintain professional autonomy, ensure patient safety, and address medico-legal concerns related to data accessibility [65]. The emphasis remains on human conversation and empathy as the core of medical practice, with AI serving as a supportive, rather than a replacement, technology [65].

### 2.6.2 Challenges and Limitations in Human-AI Collaboration in Clinical Documentation

Despite the numerous benefits of human-AI collaboration in documentation, several challenges and limitations need to be addressed. One major concern is the potential for AI systems to perpetuate or amplify existing biases present in the data on which they are trained [77]. This can lead to unfair or discriminatory outcomes, particularly in areas such as healthcare and law enforcement [77]. In the medical field, AI systems might inadvertently learn and replicate biases present in historical medical data, potentially leading to disparities in care [107]. The development of robust methods for detecting and mitigating bias in AI systems is necessary for ensuring equitable and fair outcomes [107]. Model cards, which provide detailed information about a model’s performance characteristics and potential biases, are one approach to increasing transparency and accountability [77].

The integration of AI into existing workflows can be disruptive and require significant changes in organizational processes and practices [12]. Careful planning and stakeholder engagement are necessary to ensure a smooth transition and minimize disruption [12]. Concerns about job displacement due to AI automation also need to be addressed through careful workforce planning and reskilling initiatives [12].

The issue of data quality and availability is also a significant limitation. AI systems require large amounts of high-quality data for training, and the lack of readily available data can hinder the development and deployment of AI-powered documentation systems [78]. Data privacy and security are also important considerations, particularly in healthcare and other sensitive domains [78]. Robust data governance frameworks are needed to ensure the ethical and responsible use of AI in documentation [78]. The complexity of AI systems can make it challenging to identify and address errors or malfunctions [36]. Rigorous testing and validation are essential to ensure the safety and reliability of AI-powered documentation systems [36].

# Chapter 3

## Methodology

This chapter outlines the methodological approach used to assess the quality of SOAP notes in pediatric rehabilitation. It describes the study population and details the data-collection process, including the transformation of scratch notes into structured SOAP notes. The chapter explains the anonymization process implemented to ensure unbiased evaluations, followed by an overview of the human-editing procedures applied to AI-generated SOAP notes. The rubric criteria and measurement framework used to evaluate documentation quality are then presented, along with a description of the evaluation process designed to ensure consistency and reliability in assessments.

### 3.1 Study Population

The study population consisted of a comprehensive set of SOAP notes obtained from the KidsAbility Centre for Child Development, a pediatric occupational therapy clinic in Ontario. A total of 432 SOAP notes were analyzed, divided into four distinct pools, each containing 108 notes. The first pool (K) comprised human-authored SOAP notes from occupational therapists. The second pool (E) consisted of SOAP notes generated using Copilot, an AI-assisted tool, which an occupational therapist subsequently edited. The third pool (U) included SOAP notes generated by KAUWbot, an off-the-shelf fine-tuned LLM developed for KidsAbility, which were left unedited, representing raw AI output. The fourth pool (T) featured KAUWbot-generated SOAP notes that an occupational therapist edited.

Table 3.1: The Four SOAP Note Pools and Their Description

| Note Pools | Number | Description                                  |
|------------|--------|----------------------------------------------|
| K          | 108    | Human-authored SOAP notes                    |
| E          | 108    | Copilot-generated SOAP notes (edited by OTs) |
| U          | 108    | KAUWbot-generated SOAP notes (unedited)      |
| T          | 108    | KAUWbot-generated SOAP notes (edited by OTs) |

## 3.2 Data Collection

Rehabilitation in pediatric occupational therapy is inherently subjective, as it revolves around identifying and addressing the unique functional goals that each child aspires to achieve but finds challenging [111]. Occupational therapists work collaboratively with children and their families to develop individualized treatment plans tailored to improve specific areas of function, such as motor skills, sensory processing, self-care, and social participation [111].

To support this research, KidsAbility collected 432 SOAP notes from its EHR system. These notes were prepared between January 2023 and June 2024 to ensure a comprehensive and representative sample of clinical documentation practices during this period. The dataset was evenly distributed across the four key pools.

To safeguard patient privacy, all collected SOAP notes underwent rigorous anonymization. Personal identifiers, such as names, dates of birth, and addresses, as well as any other information that could potentially reveal patient identities, were systematically removed before analysis.

## 3.3 SOAP Note Anonymization Process

The anonymization of SOAP notes is crucial for maintaining the integrity of the evaluation process and ensuring evaluators’ unbiased assessments. The anonymization is carried out in two key steps, each designed to conceal the origin of the notes while allowing for systematic analysis.

### 3.3.1 Step 1: Assigning Letters to Each Pool

The anonymization process began by assigning specific letter codes to each pool of SOAP notes to distinguish their sources while maintaining evaluator blinding. Human-authored notes were labeled “K,” Copilot AI-generated and OT-edited notes were marked “E,” unedited KAUWbot-generated notes were assigned “U,” and edited KAUWbot notes were designated “T.”

### 3.3.2 Step 2: Assigning Numbers to the Letters

To enhance anonymization, a second layer of randomization was applied by incorporating numbers into the letter-assigned codes. This process ensured that evaluators could not infer a note’s source based on the letter alone. Each SOAP note was assigned a unique alphanumeric code (e.g., 6N60, N444, 94N3), effectively concealing its origin and maintaining evaluator blinding.

| Sample Prompt for Copilot                                                                                                                                    | Sample Scratch Note                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Generated Copilot SOAP Note                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| I am an OT in a school board supporting a student. Turn my scratch notes below into a SOAP note in bullet point form. Do not use any identifying information | <p>The OT consulted with school team. OT followed up with caregiver on telephone after visit. School team reported the following:</p> <ul style="list-style-type: none"> <li>• Student has difficulties with drinking from cup at school but not home</li> <li>• Will throw cup at times</li> <li>• Requests hands over hand support</li> <li>• Looking to increase independence</li> </ul> <p>Caregiver reported the following:</p> <ul style="list-style-type: none"> <li>• purchased drinking cup</li> <li>• Drinks at home out of cup independently, practiced 1-2 times</li> <li>• OT observed student during lunch, student seated in PAL chair</li> <li>• OT F/U with family regarding drinking cup to support oral motor movements and development of using of straw.</li> <li>• Student would benefit from practice with drinking cup over summer</li> <li>• support development of self-care task of independent drinking with cup</li> <li>• OT to follow up with summer visit</li> </ul> | <p><b>Subjective</b></p> <ul style="list-style-type: none"> <li>- The school team reports the student has difficulties drinking from a cup at school but not at home.</li> <li>- At times, the student throws the cup and requests hand-over-hand support.</li> <li>- The team is working on increasing the student's independence with drinking.</li> </ul> <p><b>Objective</b></p> <ul style="list-style-type: none"> <li>- Caregiver reports purchasing a new drinking cup, and the student drinks independently at home.</li> <li>- The student has practiced drinking at home 1-2 times.</li> <li>- OT observed the student seated in a PAL chair during lunch at school.</li> <li>- Follow-up with the family was conducted to discuss using the new drinking cup to support oral motor development and straw use.</li> </ul> <p><b>Assessment</b></p> <ul style="list-style-type: none"> <li>- The student demonstrates challenges with independent drinking at school, showing a difference in behavior compared to home.</li> <li>- The student would benefit from increased practice using the new cup to support oral motor skills and develop independence in drinking.</li> </ul> <p><b>Plan</b></p> <ul style="list-style-type: none"> <li>- Recommend continued practice with the new drinking cup over the summer to further develop the self-care task of independent drinking.</li> <li>- OT to follow up with a summer visit to assess progress and understand the differences between the home and school environments.</li> </ul> |

Figure 3.1: Sample Prompt for Copilot, Sample Scratch Note and Generated Copilot SOAP Note [5].

### 3.4 From Scratch Notes to SOAP Notes

Occupational therapists often rely on the creation of scratch notes as part of their clinical documentation process [53]. Scratch notes are succinct, bullet-point entries that serve as quick, informal records of the most relevant aspects of a therapy session. These notes, typically composed during or immediately after therapy sessions, focus on capturing key observations, patient responses, therapeutic interventions, and any noteworthy events that occur during treatment. Given the fast-paced and dynamic nature of clinical environments, especially in pediatric rehabilitation, these scratch notes are designed to be brief yet comprehensive enough to ensure that critical details are not overlooked [83].

The primary advantage of scratch notes lies in their ability to be written rapidly without interrupting the flow of therapy, thereby allowing OTs to maintain their focus on patient interaction while simultaneously documenting essential information [83]. Following the completion of therapy sessions, OTs in this research transition from these shorthand scratch notes to more structured and detailed documentation in the form of SOAP notes. OTs utilize AI-driven systems such as Copilot and KAUWbot (depicted in Figure 3.1) to facilitate this transition.

## 3.5 Human Editing Process of SOAP Notes

The human editing process plays a critical role in ensuring the accuracy, relevance, and completeness of SOAP notes generated by AI tools such as Copilot and KAUWbot. Following the generation of an initial draft by these AI systems, KidsAbility OTs are tasked with reviewing the content and verifying its alignment with clinical standards.

During the editing process, OTs scrutinize the content of each section—Subjective, Objective, Assessment, and Plan—to ensure that the information is correctly interpreted and organized. The OTs’ primary responsibility is to confirm that the AI has accurately transcribed patient observations, therapeutic interventions, and clinical assessments. In cases where the AI has misinterpreted data or introduced inaccuracies, the OTs make the necessary corrections to maintain the integrity of the clinical record. These corrections can range from fixing errors in the AI’s understanding of the input data to addressing misrepresentations of patient responses or therapeutic actions.

A common challenge encountered during the editing process is the improper categorization of information within the SOAP note. Although AI tools like Copilot and KAUWbot are sophisticated in processing and structuring clinical data, they sometimes misplace or incorrectly assign details. For example, subjective observations—such as a patient’s self-reported symptoms or feelings—may be mistakenly placed in the objective section, or assessment findings may be wrongly categorized under the plan. According to previous studies, such misclassifications are not uncommon in AI-generated clinical documentation [41].

Beyond correcting factual and structural errors, OTs also personalize the AI-generated notes to reflect their clinical style and preferences. Each OT has a unique approach to documenting patient care, shaped by their professional experience, institutional guidelines, and personal preferences. As such, the editing process often includes reformatting the text, adjusting the tone, and incorporating specific clinical terminology that the AI might not have included.

## 3.6 Rubric Criteria and Measurements

The quality of SOAP notes in this research was evaluated in a blinded manner using a rubric based on five specific criteria, which were adapted from the Physician Documentation Quality Instrument (PDQI-9) [102]. The PDQI-9 is a condensed version of the original Physician Documentation Quality Instrument [103], designed to assess clinical documentation quality across nine key dimensions: *completeness*, *correctness*, *appropriateness*, *organization*, *clarity*, *conciseness*, *comprehensiveness*, *usefulness*, and *information synthesis*. These dimensions provide a comprehensive framework for evaluating the effectiveness of clinical notes, with an emphasis on improving patient care outcomes. This research uses the PDQI-9 to assess the reliability and consistency of EHR documentation, highlighting the importance of structured, accurate entries in modern healthcare environments [102].

While the PDQI-9 offers a thorough evaluation across nine dimensions, the current research focused on a more custom set of five criteria—***Clear***, ***Complete***, ***Concise***, ***Relevant***, and ***Organized***—for several key reasons. These five criteria were selected because they closely align with the structured nature of SOAP notes, which prioritize simplicity and clarity to maximize clinical efficiency. SOAP notes are designed to deliver essential patient information clearly and concisely, enabling healthcare providers to quickly grasp critical details without unnecessary complexity [86]. The chosen criteria reflect these priorities and are integral to ensuring effective communication in SOAP-note format.

For instance, *clarity* ensures that the note is easily understandable by other healthcare professionals, facilitating efficient collaboration and continuity of care. *Completeness* ensures that all relevant information is documented without omissions, while *relevance* ensures that only pertinent details are included, eliminating unnecessary or extraneous content. *Conciseness* emphasizes the importance of brevity, reducing redundancy and enhancing the note’s efficiency, while *organization* ensures that the information flows logically and adheres to the standardized SOAP structure, promoting ease of use and interpretation.

The evaluation instrument employed in this research assessed these components, each rated on a three-point Likert scale. A score of one represented low quality, two indicated moderate quality, and three denoted high quality. The total score achievable for each SOAP note ranged from a minimum of five to a maximum of 15.

### 3.7 The Evaluation Process

The quality of SOAP notes generated through both human-only and AI-assisted methods was assessed by four independent evaluators from KidsAbility, each an experienced clinician. To ensure objectivity and minimize potential bias, the evaluators’ identities were kept confidential throughout the research. This approach was designed to prevent any external influences from shaping the evaluation outcomes, thereby ensuring the integrity and reliability of the assessments. Each evaluator reviewed the full set of SOAP notes, meaning that each note was evaluated four times—once by each of the four clinicians. Given the total dataset of 432 SOAP notes, drawn from four distinct pools, the study resulted in 1,728 individual evaluations.

Each SOAP note was paired with an evaluation instrument (Figure 3.2) that included the custom rubric criteria essential to the research. This custom evaluation tool ensured consistency across all assessments by providing a clear, structured framework for the evaluators.

## Clinician Note Quality Evaluation Instrument

| Criterion        | Description / Requirement                                     | Rating Scale                                                                                                                     |
|------------------|---------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| <b>Clear</b>     | The note is easily understandable.                            | <input type="radio"/> 1. Not clear<br><input type="radio"/> 2. Partially clear<br><input type="radio"/> 3. Very clear            |
| <b>Complete</b>  | The note fully reports on the session.                        | <input type="radio"/> 1. Incomplete<br><input type="radio"/> 2. Partially complete<br><input type="radio"/> 3. Fully complete    |
| <b>Concise</b>   | The note is focused and brief.                                | <input type="radio"/> 1. Not concise<br><input type="radio"/> 2. Moderately concise<br><input type="radio"/> 3. Very concise     |
| <b>Relevant</b>  | The content of the note directly contributes to patient care. | <input type="radio"/> 1. Irrelevant<br><input type="radio"/> 2. Partially relevant<br><input type="radio"/> 3. Relevant          |
| <b>Organized</b> | Information is properly grouped and structured.               | <input type="radio"/> 1. Disorganized<br><input type="radio"/> 2. Partially organized<br><input type="radio"/> 3. Well organized |

**Total Score:** \_\_\_\_\_ / 15

---

### Instructions for Evaluators:

- **Clear:** Assess if the note is readily understandable and without ambiguity.
- **Complete:** Check if all relevant aspects of the session are addressed.
- **Concise:** Ensure the note is direct and to the point without unnecessary detail.
- **Relevant:** Verify that the information in the note directly contributes to patient care.
- **Organized:** Look for proper grouping and structure of information.

Each criterion should be rated on a scale from 1 to 3, where 1 indicates the lowest quality, 2 indicates moderate quality, and 3 indicates the highest quality.

Figure 3.2: SOAP Note Evaluation Instrument.

# Chapter 4

## Results and Analysis

This chapter presents the results and analysis of the study, focusing on the quality assessment of SOAP notes across different pools. First, the mean scores for the four SOAP-note pools are reported to provide an overview of overall performance. This is followed by descriptive statistics summarizing the distribution of quality scores. A statistical analysis using ANOVA is then conducted to compare mean quality scores across the pools and determine significant differences. Finally, the performance of each SOAP-note pool is evaluated based on the criteria—clarity, completeness, conciseness, relevance, and organization—to gain deeper insights into specific areas of strength and limitation.

### 4.1 Mean Scores for the Four SOAP Note Pools

The average scores shown in Table 4.1 for human-authored SOAP notes (K) reveal notable differences in clinician scoring. Clinician 2 and Clinician 4 consistently scored higher, particularly in relevance (2.94) and clarity (2.59 and 2.48, respectively). Clinician 1 scored highest for conciseness (2.64), while Clinician 4 demonstrated strong performance in completeness (2.55) and organization (2.40). In contrast, Clinician 3 displayed lower average scores across all criteria, especially for completeness (1.28) and organization (1.33).

Table 4.1: Human-Authored SOAP Notes (K)

| Criterion | Clinician 1 | Clinician 2 | Clinician 3 | Clinician 4 |
|-----------|-------------|-------------|-------------|-------------|
| Clear     | 2.60        | 2.59        | 1.73        | 2.48        |
| Complete  | 2.09        | 2.27        | 1.28        | 2.55        |
| Concise   | 2.64        | 2.87        | 1.73        | 2.83        |
| Relevant  | 2.63        | 2.94        | 2.28        | 2.94        |
| Organized | 1.96        | 2.36        | 1.33        | 2.40        |

The data for Copilot-generated SOAP notes (Table 4.2) edited by occupational therapists (E) demonstrates strong performance across clinicians, with consistently high scores in clarity, completeness, relevance, and organization. Clinician 2 stands out with the

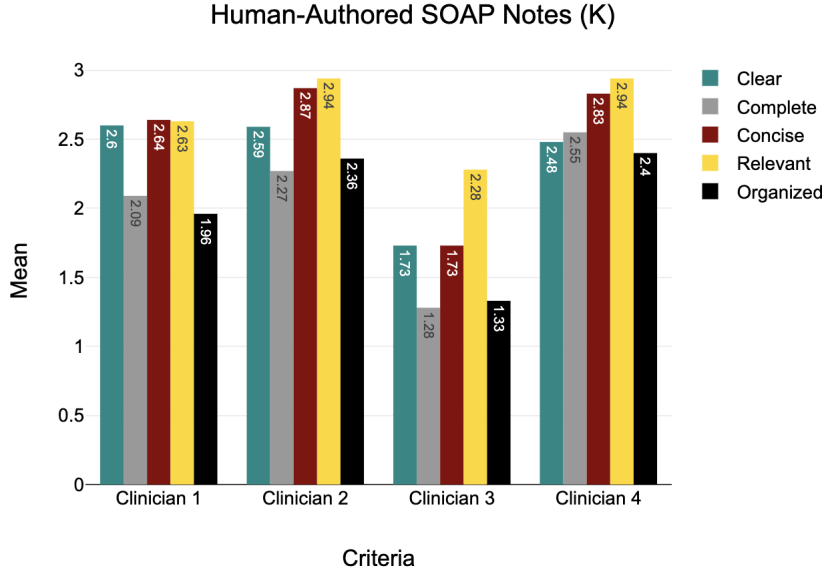


Figure 4.1: Histogram of Clinician Scoring of Human-Authored SOAP Notes.

highest scores in conciseness (2.81), relevance (2.90), and organization (2.73). Clinician 4 also scored high, particularly in completeness (2.85) and organization (2.76). Clinician 3, while scoring lower overall, still demonstrated a relatively strong score in relevance (2.53) compared to other criteria.

Table 4.2: Copilot SOAP Notes - OT edited (E)

| Criterion | Clinician 1 | Clinician 2 | Clinician 3 | Clinician 4 |
|-----------|-------------|-------------|-------------|-------------|
| Clear     | 2.73        | 2.76        | 2.33        | 2.68        |
| Complete  | 2.67        | 2.74        | 2.25        | 2.85        |
| Concise   | 2.16        | 2.81        | 2.13        | 2.47        |
| Relevant  | 2.82        | 2.90        | 2.53        | 2.75        |
| Organized | 2.19        | 2.73        | 2.24        | 2.76        |

The analysis of unedited KAUWbot SOAP notes (U) in Table 4.3 indicates generally strong scores in relevance and clarity, with Clinician 2 providing the highest scores in relevance (2.97) and conciseness (2.92). Clinician 1 also provided a high score in clarity (2.76) and completeness (2.68). In contrast, Clinician 3 consistently scored lower, particularly in completeness (1.98) and organization (2.23). Organization scores were relatively lower across all clinicians.

The data for KAUWbot SOAP notes (Table 4.4) edited by occupational therapists (T) shows strong overall performance, particularly in relevance, where Clinician 2 achieved the highest score (2.99) across all criteria. Clinician 1 also excelled in clarity (2.85), while Clinician 4 demonstrated solid scores in completeness (2.79) and conciseness (2.69). Clinician 3's scores were consistently lower, especially in completeness (2.09) and organization (2.31). Organization scores remained relatively weaker across clinicians. Overall, the OT-edited KAUWbot notes exhibit notable improvements in clarity, relevance, and completeness.

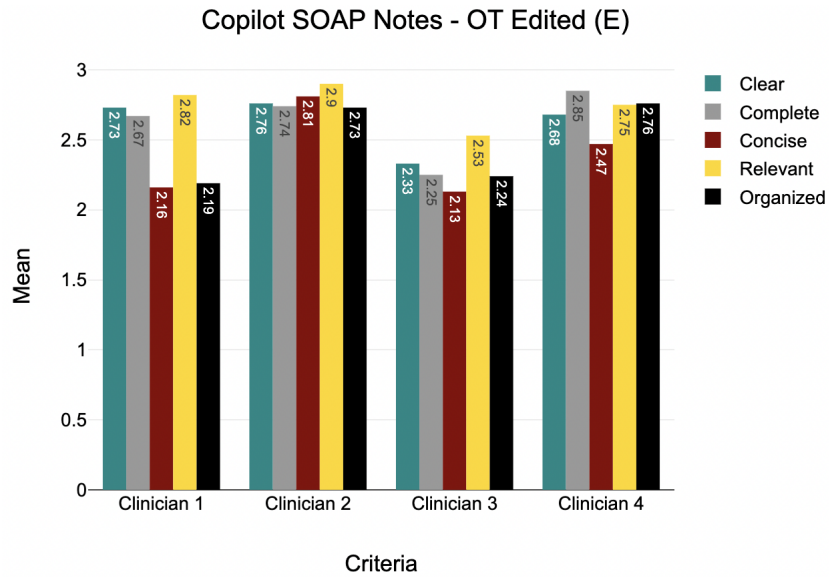


Figure 4.2: Histogram of Clinician Scoring of Copilot-Generated SOAP Notes Edited by OTs.

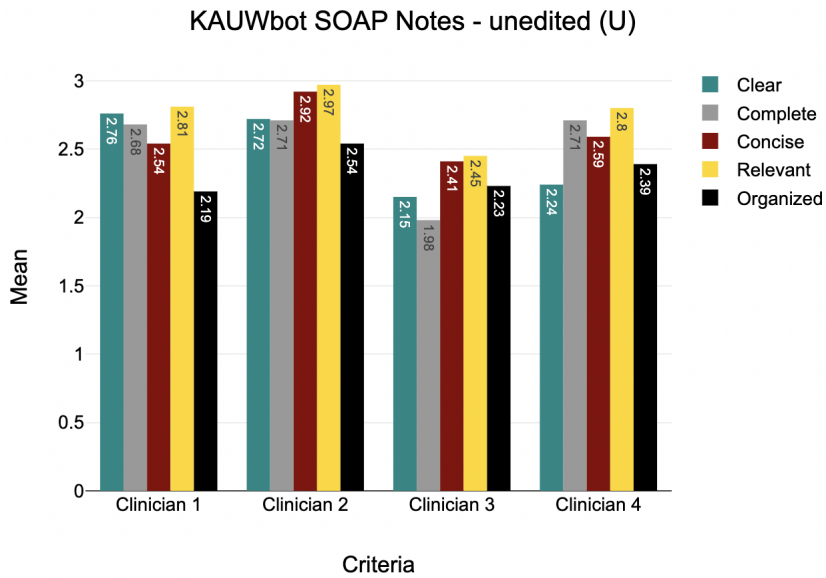


Figure 4.3: Histogram of Clinician Scoring of KAUWbot-Generated SOAP Notes Unedited.

Table 4.3: KAUWbot SOAP Notes - unedited (U)

| <b>Criterion</b> | <b>Clinician 1</b> | <b>Clinician 2</b> | <b>Clinician 3</b> | <b>Clinician 4</b> |
|------------------|--------------------|--------------------|--------------------|--------------------|
| Clear            | 2.76               | 2.72               | 2.15               | 2.24               |
| Complete         | 2.68               | 2.71               | 1.98               | 2.71               |
| Concise          | 2.54               | 2.92               | 2.41               | 2.59               |
| Relevant         | 2.81               | 2.97               | 2.45               | 2.80               |
| Organized        | 2.19               | 2.54               | 2.23               | 2.39               |

Table 4.4: KAUWbot SOAP Notes - OT edited (T)

| <b>Criterion</b> | <b>Clinician 1</b> | <b>Clinician 2</b> | <b>Clinician 3</b> | <b>Clinician 4</b> |
|------------------|--------------------|--------------------|--------------------|--------------------|
| Clear            | 2.85               | 2.81               | 2.32               | 2.43               |
| Complete         | 2.69               | 2.83               | 2.09               | 2.79               |
| Concise          | 2.52               | 2.90               | 2.46               | 2.69               |
| Relevant         | 2.90               | 2.99               | 2.51               | 2.83               |
| Organized        | 2.27               | 2.74               | 2.31               | 2.57               |

## 4.2 Quality Performance for the Four SOAP-Note Pools

### 4.2.1 Descriptive Statistics of Quality Scores

The descriptive statistics provided for the evaluators’ quality scores across the four SOAP-note pools—human-authored (K), Copilot OT-edited (E), KAUWbot unedited (U), and KAUWbot OT-edited (T)—highlight differences in overall quality and variability.

Table 4.5: Summary of Mean Values for Evaluators’ (Clinician) Quality Scores for the SOAP-Note Pools

| <b>Evaluators</b> | <b>K</b> | <b>E</b> | <b>U</b> | <b>T</b> |
|-------------------|----------|----------|----------|----------|
| Clinician 1       | 11.80    | 12.57    | 12.97    | 13.21    |
| Clinician 2       | 12.92    | 13.94    | 13.73    | 14.27    |
| Clinician 3       | 8.19     | 11.44    | 11.20    | 11.70    |
| Clinician 4       | 12.06    | 13.38    | 12.73    | 13.32    |

From Table 4.5, it is evident that Clinician 2 consistently provided the highest scores across all note pools, ranging from 12.92 for the K pool to 14.27 for the T pool. This suggests that Clinician 2 either perceives higher quality in the notes or applies more lenient evaluation criteria than others. Clinician 3 provided the lowest ratings across all pools, with scores ranging from 8.19 for the K pool to 11.70 for the T pool, indicating a more critical or stricter approach to evaluation. The highest mean scores overall were given to the T pool (KAUWbot OT-edited) by Clinician 2 (14.27), reflecting the perceived improvement in quality when AI-generated notes were edited by OTs. The lowest mean rating was assigned to the K pool (human-authored) by Clinician 3 (8.19).

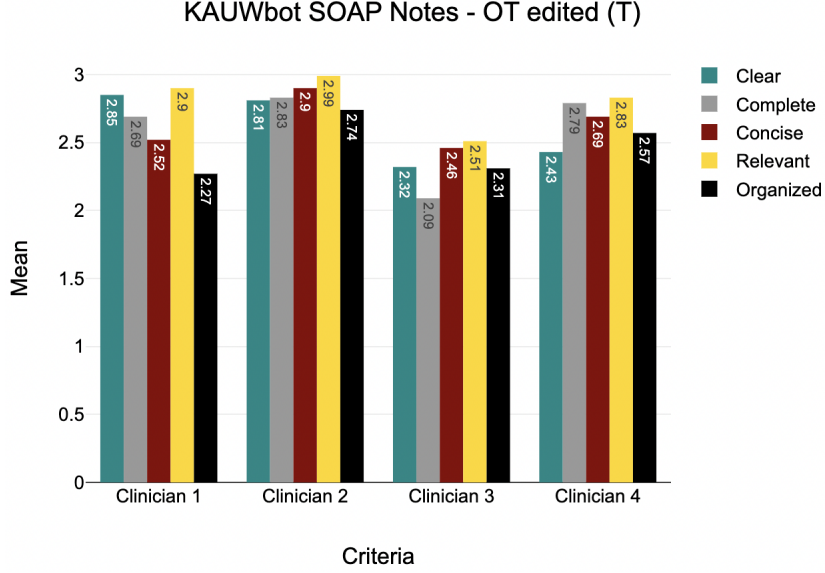


Figure 4.4: Histogram of Clinician Scoring of KAUWbot-Generated SOAP Notes Edited by OTs.

Table 4.6: Summary of Data for Evaluators' (Clinician) Quality Scores for the SOAP-Note Pools

| Note Pools | Mean  | Median | Std. Dev. |
|------------|-------|--------|-----------|
| K          | 11.49 | 12.36  | 2.28      |
| E          | 12.83 | 12.98  | 1.09      |
| U          | 12.66 | 12.85  | 1.06      |
| T          | 13.12 | 13.26  | 1.06      |

The summary statistics in Table 4.6 provide further insight. Mean scores progressively increase across the note pools: K (11.49), E (12.83), U (12.66), and T (13.12). This progression highlights that AI-generated notes and subsequent OT editing are associated with higher perceived quality compared to human-authored notes. The *median* ratings align closely with the means, indicating relatively symmetric distributions of scores within each pool. However, variability differs across pools, as shown by the standard deviation values. The K pool has the highest standard deviation (2.28), suggesting significant variability in the perceived quality of human-authored notes across evaluators. The E, U, and T pools exhibit lower and nearly identical standard deviations (ranging from 1.06 to 1.09), reflecting greater consistency in quality ratings for AI-assisted or edited notes.

When comparing the note pools, the T pool (KAUWbot OT-edited) consistently achieved the highest ratings, with a mean of 13.12 and a median of 13.26. This indicates that OT editing further enhances the quality of AI-generated notes, making them the most refined. The K pool (human-authored) received the lowest mean rating (11.49) and showed the greatest variability. Notably, the unedited KAUWbot notes (U) outperformed the human-authored notes (K), with a higher mean rating (12.66 vs. 11.49) and lower variability (standard deviation of 1.06 vs. 2.28).

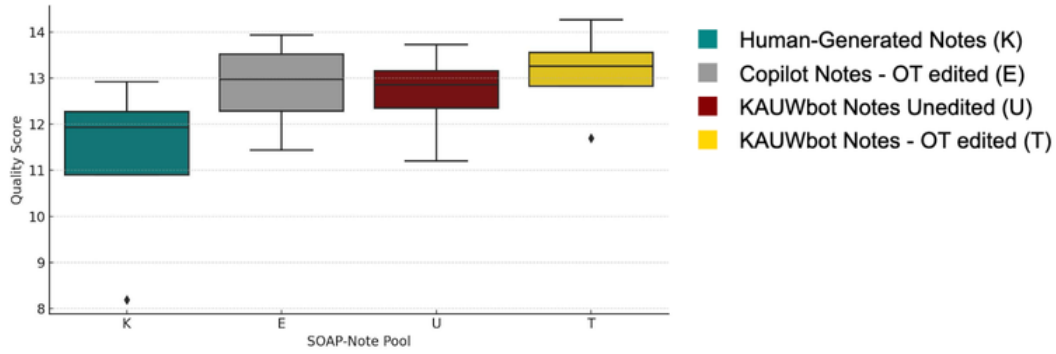


Figure 4.5: Box Plot of SOAP Note Quality Scores by Evaluators

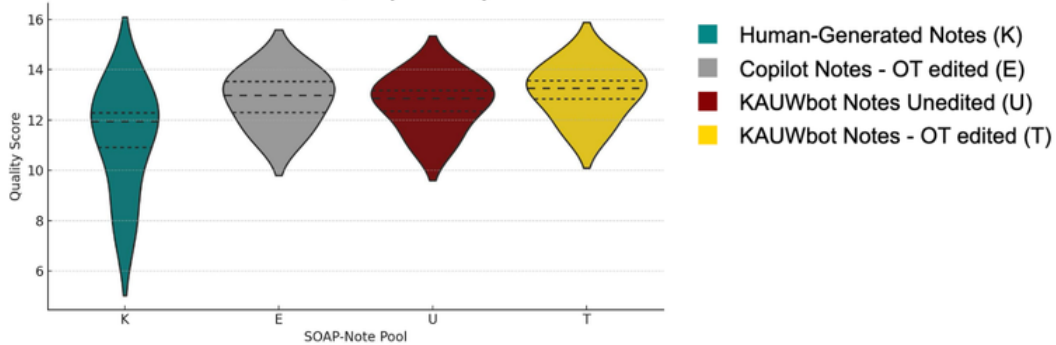


Figure 4.6: Violin Plot of SOAP Note Quality Scores by Evaluators

The box plot in Figure 4.5 reveals the distribution of quality scores across the four SOAP-note pools (K, E, U, T). Among them, Pool T (edited KAUWbot notes) has the highest median score, suggesting that it was rated most favourably by the evaluators. In contrast, Pool K (human-authored notes) has the lowest median, indicating comparatively lower ratings. The variability in scores is most visible in Pool K, reflecting inconsistencies in clinician evaluations. Pools E (edited Copilot notes) and U (unedited KAUWbot notes) exhibit relatively similar distributions, with less variation compared to K.

The violin plot in Figure 4.6 provides a deeper look at the distribution density of quality scores, showing where scores are more concentrated. Pool K (human-authored notes) has a wider distribution toward the lower end, indicating greater variability and a lower overall central tendency. Pool T (edited KAUWbot notes) demonstrates a denser concentration of scores in the upper range, reinforcing its higher ratings across evaluators. Pools E (edited Copilot notes) and U (unedited KAUWbot notes) show a balanced and relatively symmetrical spread, suggesting consistency in how evaluators assessed these SOAP notes. The violin plot further highlights that while some pools, such as T, received consistently high scores, others, like K, experienced greater variation, potentially indicating differences in clinician perspectives.

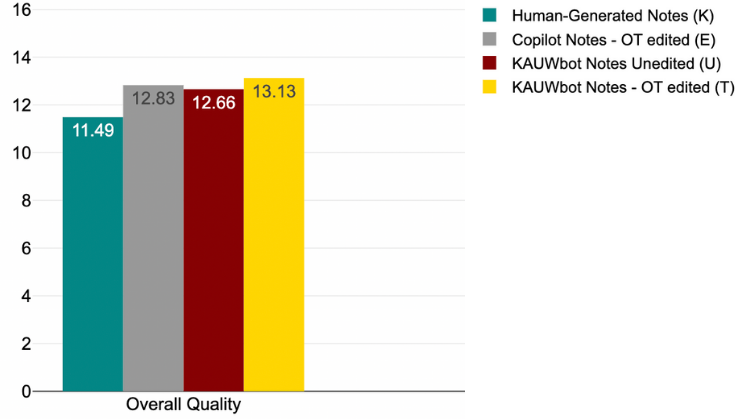


Figure 4.7: Histogram of Mean Quality Scores for SOAP-Note Pools [5].

### 4.2.2 ANOVA to Compare the Mean Quality Scores

Analyzing the differences in quality scores across various SOAP-note pools in this research necessitates a robust statistical method capable of comparing group means. One such method is the Analysis of Variance (ANOVA), which is particularly suitable for this research because it allows the researcher to determine whether there are statistically significant differences in the quality scores assigned to the four distinct SOAP-note pools. ANOVA is a widely used statistical technique in social sciences and healthcare research for comparing multiple groups simultaneously, without inflating the Type I error rate (i.e., rejecting a true null hypothesis) [26].

The dependent variable in this research is the quality score assigned by evaluators, and the independent variable is the SOAP-note pool (K, E, U, T). A one-way ANOVA can test whether the mean quality scores differ significantly between the note pools, accounting for variability among evaluators [114].

Table 4.7: Summary of Analysis of Variance (ANOVA) with  $\alpha = 0.05$

| Source         | DF | SS    | MS   | $F - stat$ | $p - value$ | $F - crit$ |
|----------------|----|-------|------|------------|-------------|------------|
| Between Groups | 3  | 6.17  | 2.06 | 0.96       | 0.45        | 3.49       |
| Within Groups  | 12 | 25.83 | 2.15 |            |             |            |
| Total          | 15 | 32    |      |            |             |            |

The ANOVA Table 4.7 summarizes the results of the analysis conducted to assess whether there are statistically significant differences in the quality scores among the four SOAP-note pools. The table includes components such as the degrees of freedom (DF), sums of squares (SS), mean squares (MS),  $F - statistic$ ,  $p - value$ , and the critical F-value ( $F - crit$ ).

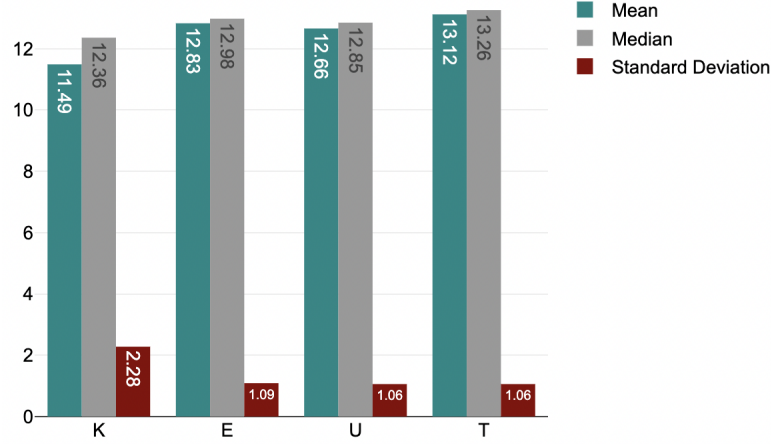


Figure 4.8: Histogram of Mean, Median, and Standard Deviation of Quality Scores for SOAP-Note Pools.

### Degrees of Freedom (DF):

The degrees of freedom for the “Between Groups” source is 3, corresponding to the four-note pools minus one (number of groups - 1). For “Within Groups,” the degree of freedom is 12, calculated as the total number of observations minus the number of groups (16 - 4). The total degree of freedom is 15, which is the sum of the between-group and within-group degrees of freedom. These degrees of freedom are crucial for determining the  $F$  - *statistic* and its associated  $p$  - *value*.

$DF_{\text{Between Groups}}$ : This is the number of groups minus 1:

$$DF_{\text{Between Groups}} = k - 1 = 4 - 1 = 3$$

$DF_{\text{Within Groups}}$ : This is the total number of observations minus the number of groups:

$$DF_{\text{Within Groups}} = N - k = 16 - 4 = 12$$

$DF_{\text{Total}}$ : This is the total number of observations minus 1:

$$DF_{\text{Total}} = N - 1 = 16 - 1 = 15$$

### Sums of Squares (SS):

The total variability in the data is represented by the total SS, which is 32. This total variability is partitioned into two components: the variability between groups (SS Between = 6.17) and the variability within groups (SS Within = 25.83). The relatively low SS Between compared to SS Within suggests that a substantial portion of the total variability is attributable to differences within groups rather than differences between the SOAP note pools.

### Mean Squares (MS):

The mean squares are calculated by dividing the sums of squares by their respective degrees of freedom. For the “Between Groups” source, the MS is 2.06 ( $SS_{\text{Between Groups}} \div DF_{\text{Between Groups}}$ ), and for the “Within Groups,” the MS is 2.15 ( $SS_{\text{Within Groups}} \div DF_{\text{Within Groups}}$ ). These mean square values are used to compute the  $F$  – *statistic*, which quantifies the ratio of between-group variability to within-group variability.

$$MS_{\text{Between Groups}} = \frac{SS_{\text{Between Groups}}}{DF_{\text{Between Groups}}} = \frac{6.17}{3} = 2.06$$
$$MS_{\text{Within Groups}} = \frac{SS_{\text{Within Groups}}}{DF_{\text{Within Groups}}} = \frac{25.83}{12} = 2.15$$

### $F$ – *statistic* and $p$ – *value*:

The calculated  $F$  – *statistic* is 0.96, derived by dividing the MS Between (2.06) by the MS Within (2.15). The  $p$  – *value* associated with this  $F$  – *statistic* is 0.45, which is much greater than the conventional significance level of 0.05. This indicates no statistically significant difference in the mean quality ratings among the four note pools. In other words, the variability observed between the groups is not sufficiently large to conclude that the note pools differ in quality.

$$F = \frac{MS_{\text{Between Groups}}}{MS_{\text{Within Groups}}} = \frac{2.06}{2.15} \approx 0.96$$

### F-Critical Value (F-crit):

The critical F-value at a 0.05 significance level, given 3 and 12 degrees of freedom, is 3.49. Since the calculated  $F$  – *statistic* (0.96) is less than the  $F$  – *critical* value (3.49), we fail to reject the null hypothesis. This further confirms that the differences in mean quality ratings among the four SOAP-note pools are not statistically significant.

## 4.3 Clarity Performance for the Four SOAP-Note Pools

The analysis of clarity scores for the four SOAP-note pools reveals important insights into the evaluators’ assessments. The two tables (Table 4.8 and Table 4.9) summarize the mean scores for each evaluator across the note pools, as well as aggregate descriptive statistics, including mean, median, and standard deviation.

The clarity scores assigned by Clinician 1 show a progressive increase across the note pools, with the lowest score for K (2.60) and the highest for T (2.85). Clinician 2 also followed a similar trend, with scores ranging from 2.59 for K to 2.81 for T. These patterns

Table 4.8: Summary of Mean Values for Evaluators’ (Clinician) Clarity Scores for the SOAP-Note Pools

| <b>Evaluators</b> | <b>K</b> | <b>E</b> | <b>U</b> | <b>T</b> |
|-------------------|----------|----------|----------|----------|
| Clinician 1       | 2.60     | 2.73     | 2.76     | 2.85     |
| Clinician 2       | 2.59     | 2.76     | 2.72     | 2.81     |
| Clinician 3       | 1.73     | 2.33     | 2.15     | 2.32     |
| Clinician 4       | 2.48     | 2.68     | 2.24     | 2.43     |

Table 4.9: Summary of Data for Evaluators’ (Clinician) Clarity Scores for the SOAP-Note Pools

| <b>Note Pools</b> | <b>Mean</b> | <b>Median</b> | <b>Std. Dev.</b> |
|-------------------|-------------|---------------|------------------|
| K                 | 2.35        | 2.54          | 0.42             |
| E                 | 2.63        | 2.71          | 0.20             |
| U                 | 2.47        | 2.48          | 0.32             |
| T                 | 2.60        | 2.62          | 0.27             |

suggest a positive perception of clarity in the edited and AI-generated notes compared to the human-authored notes. In contrast, Clinician 3’s scores, while also increasing, start from a significantly lower baseline (1.73 for K) and reach only 2.32 for T. Clinician 4’s scores are relatively consistent, with slight variations, ranging from 2.24 for U to 2.68 for E.

The overall mean clarity scores reveal that E (Copilot OT-edited) achieved the highest mean score of 2.63, followed closely by T (KAUWbot OT-edited) at 2.60, indicating that editing and refinement of AI-generated notes improve perceived clarity. The U (KAUWbot unedited) and K (human-authored) pools had slightly lower means, at 2.47 and 2.35, respectively, suggesting that unedited notes and human-authored notes may lack consistency in clarity.

The median scores further support these findings, with E achieving the highest median of 2.71 and T following at 2.62. The human-authored notes (K) had the lowest median of 2.54, reflecting some variability in clarity perceptions among evaluators. The standard deviation values indicate variability in the ratings, with K showing the highest variability (0.42) and E showing the least (0.20). This suggests that human-authored notes were assessed with greater inconsistency, potentially due to differences in writing styles or evaluator interpretations, while Copilot OT-edited notes achieved more consistent clarity.

## 4.4 Completeness Performance for the Four SOAP-Note Pools

The analysis of the evaluators’ completeness scores in Table 4.10 and Table 4.11 for the SOAP-note pools reveals several notable trends in the data across different note pools.

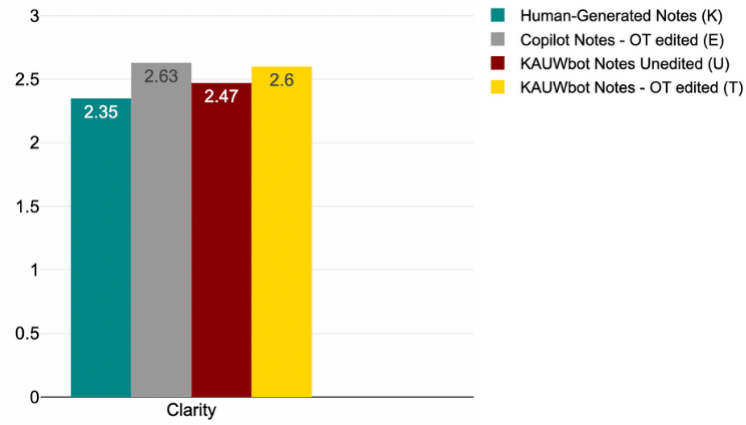


Figure 4.9: Histogram of Mean Clarity Scores for SOAP-Note Pools.

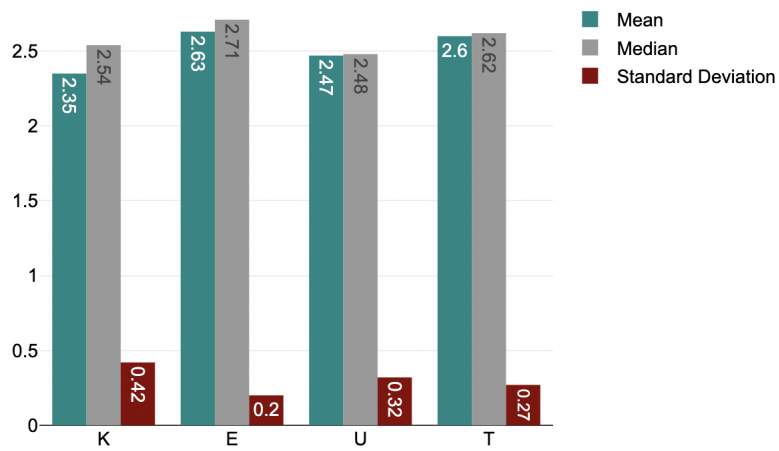


Figure 4.10: Histogram of Mean, Median, and Standard Deviation of Clarity Scores for SOAP-Note Pools.

Table 4.10: Summary of Mean Values for Evaluators’ (Clinician) Completeness Scores for the SOAP-Note Pools

| <b>Evaluators</b> | <b>K</b> | <b>E</b> | <b>U</b> | <b>T</b> |
|-------------------|----------|----------|----------|----------|
| Clinician 1       | 2.09     | 2.67     | 2.68     | 2.69     |
| Clinician 2       | 2.27     | 2.74     | 2.71     | 2.83     |
| Clinician 3       | 1.28     | 2.25     | 1.98     | 2.09     |
| Clinician 4       | 2.55     | 2.85     | 2.71     | 2.79     |

Table 4.11: Summary of Data for Evaluators’ (Clinician) Completeness Scores for the SOAP-Note Pools

| <b>Note Pools</b> | <b>Mean</b> | <b>Median</b> | <b>Std. Dev.</b> |
|-------------------|-------------|---------------|------------------|
| K                 | 2.05        | 2.18          | 0.55             |
| E                 | 2.63        | 2.71          | 0.26             |
| U                 | 2.52        | 2.70          | 0.36             |
| T                 | 2.60        | 2.74          | 0.35             |

In Table 4.10, Clinician 1 generally rates the note pools lower in terms of completeness, with a mean score of 2.09 for the K pool and higher scores for the T pool (2.69). Clinician 2 provides a similar pattern, with the highest mean score of 2.83 for the T pool, followed by scores ranging from 2.27 to 2.74 for the other pools. Clinician 3, on the other hand, rated the K pool notably lower (1.28) compared to the other pools, suggesting a possible discrepancy in their evaluation approach or a lower perceived completeness for the K pool. This trend contrasts with Clinician 4, who provided consistently higher ratings, with mean scores ranging from 2.55 for K to 2.85 for E.

The K pool in Table 4.11 has the lowest mean score (2.05) and a relatively high standard deviation of 0.55, indicating that the completeness of the human-authored SOAP notes is rated inconsistently across evaluators. The standard deviation suggests that there is a notable variability in how different evaluators perceive the completeness of the K notes. In contrast, the E pool has the highest mean score (2.63), accompanied by the lowest standard deviation (0.26), reflecting a more consistent and generally higher perception of completeness across evaluators. The U and T pools have very similar mean scores (2.52 and 2.60, respectively), with the T pool showing a slightly higher median (2.74 compared to 2.70 for U), indicating a more consistent evaluation of the KAUWbot-generated and OT-edited notes.

The differences in mean scores between the pools highlight the evaluators’ overall perception of completeness, with the Copilot OT-edited (E) pool receiving the highest and most consistent ratings, suggesting that the edits made by the OT improved the completeness of the SOAP notes. Meanwhile, the K pool (human-authored) appears to have more variability and the lowest mean rating, indicating that the human-authored notes may have been perceived as less complete or more inconsistent compared to the other pools. The KAUWbot-generated notes, both unedited (U) and OT-edited (T), received more consistent and moderately high scores, with the OT-edited pool slightly outperforming the unedited one in terms of completeness.

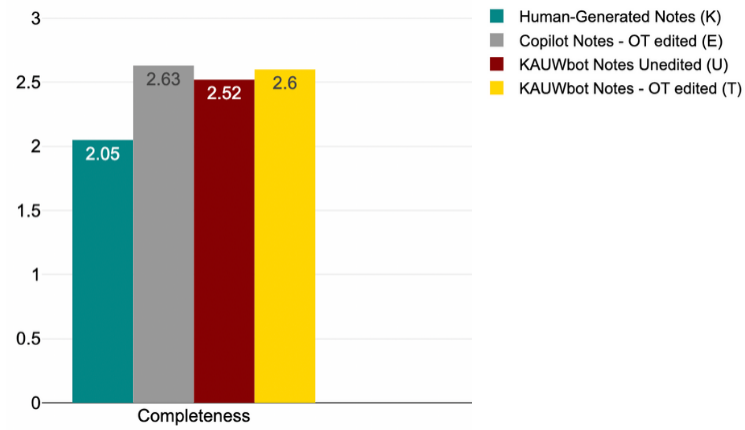


Figure 4.11: Histogram of Mean Completeness Scores for SOAP-Note Pools.

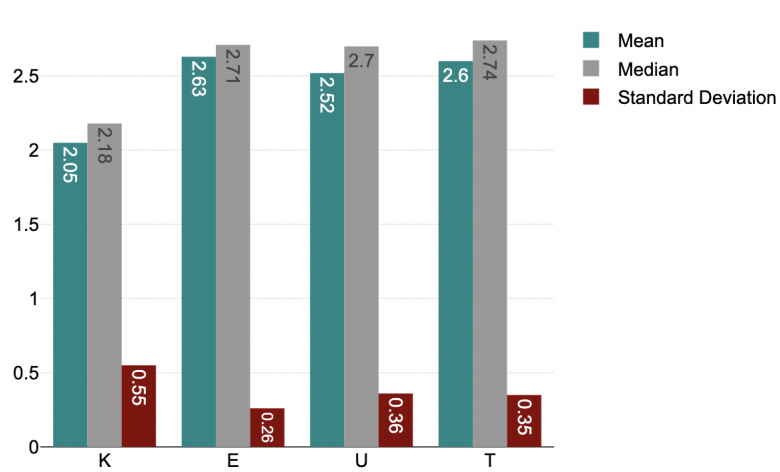


Figure 4.12: Histogram of Mean, Median, and Standard Deviation of Completeness Scores for SOAP-Note Pools.

## 4.5 Conciseness Performance for the Four SOAP Note Pools

The descriptive statistics for the evaluators’ (clinician) conciseness scores for the SOAP note pools, as shown in Table 4.12 and Table 4.13, reveal insights into how the different note pools were perceived in terms of their conciseness by the clinicians.

Table 4.12: Summary of Mean Values for Evaluators’ (Clinician) Conciseness Scores for the SOAP-Note Pools

| <b>Evaluators</b> | <b>K</b> | <b>E</b> | <b>U</b> | <b>T</b> |
|-------------------|----------|----------|----------|----------|
| Clinician 1       | 2.64     | 2.16     | 2.54     | 2.52     |
| Clinician 2       | 2.87     | 2.81     | 2.92     | 2.90     |
| Clinician 3       | 1.73     | 2.13     | 2.41     | 2.46     |
| Clinician 4       | 2.83     | 2.47     | 2.59     | 2.69     |

Table 4.13: Summary of Data for Evaluators’ (Clinician) Conciseness Scores for the SOAP-Note Pools

| <b>Note Pools</b> | <b>Mean</b> | <b>Median</b> | <b>Std. Dev.</b> |
|-------------------|-------------|---------------|------------------|
| K                 | 2.52        | 2.74          | 0.54             |
| E                 | 2.39        | 2.32          | 0.32             |
| U                 | 2.62        | 2.57          | 0.22             |
| T                 | 2.64        | 2.61          | 0.20             |

The mean conciseness scores across the four note pools vary slightly but show some notable differences in the ratings. Clinician 1 rated the conciseness of the K note pool the highest (mean = 2.64) compared to the other pools, with the lowest score for the E note pool (mean = 2.16). Clinician 2, on the other hand, rated the E note pool highest (mean = 2.81), followed by the T pool (mean = 2.90). This pattern is somewhat reversed for Clinician 3, who rated the K note pool the lowest (mean = 1.73) and the T pool highest (mean = 2.46), indicating varying clinician preferences or interpretations of conciseness in the SOAP notes. Clinician 4, in contrast, showed more consistent ratings across the note pools, with the highest mean score for the K pool (mean = 2.83), and relatively similar ratings for the other three pools (mean = 2.47 for E, 2.59 for U, and 2.69 for T).

From Table 4.13, we observe that the overall mean conciseness scores for the note pools suggest that the KAUWbot-unedited (U) and KAUWbot OT-edited (T) note pools are perceived as the most concise by the evaluators, followed closely by the human-authored (K) pool. The Copilot OT-edited (E) pool was rated the lowest in terms of conciseness, suggesting that it may have been perceived as less succinct compared to the others.

The standard deviation values reflect the variability of the conciseness scores within each note pool. The highest standard deviation is observed in the K pool (0.54), indicating that the conciseness ratings for this pool were more spread out across evaluators. The standard deviation for the E pool is lower (0.32), suggesting that evaluators generally agreed more on

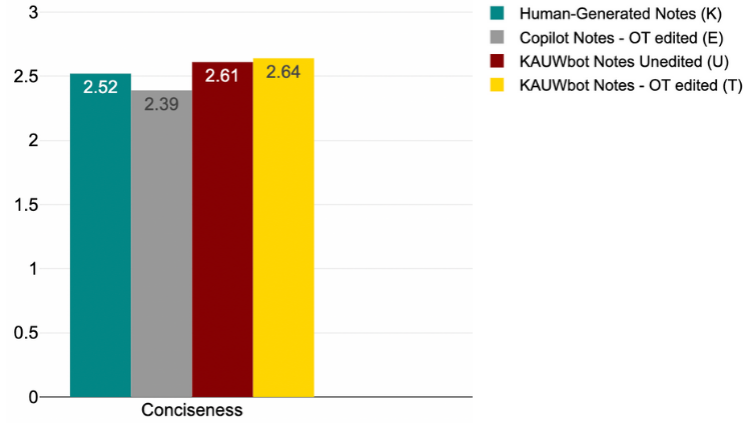


Figure 4.13: Histogram of Mean Conciseness Scores for SOAP-Note Pools.

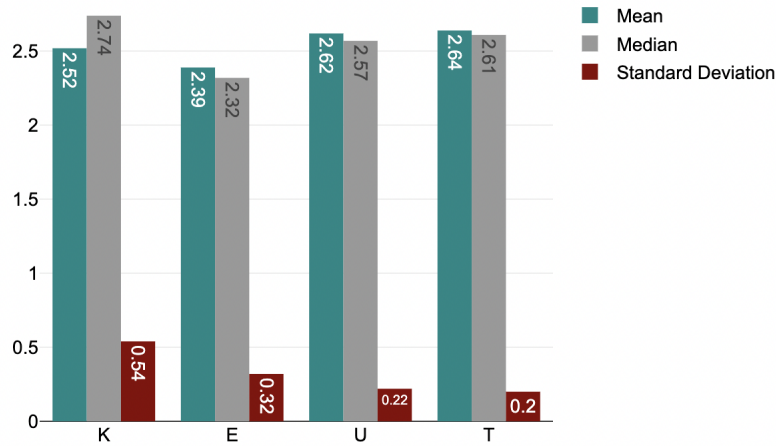


Figure 4.14: Histogram of Mean, Median, and Standard Deviation of Conciseness Scores for SOAP-Note Pools.

the conciseness of the Copilot-edited notes. The U and T pools both have lower standard deviations (0.22 and 0.20, respectively), indicating relatively high consistency in how these pools were rated for conciseness.

## 4.6 Relevance Performance for the Four SOAP-Note Pools

The descriptive statistics for the evaluators' (clinician) relevance scores for the SOAP-note pools, as shown in Table 4.14 and Table 4.15, reveal insights into how the different note pools were perceived in terms of their relevance by the clinicians.

Clinician 1 in Table 4.15 rated pool K the lowest (2.63) and pool T the highest (2.90), with scores for pools E and U close to that of T, reflecting overall consistency in their

Table 4.14: Summary of Mean Values for Evaluators' (Clinician) Relevance Scores for the SOAP-Note Pools

| <b>Evaluators</b> | <b>K</b> | <b>E</b> | <b>U</b> | <b>T</b> |
|-------------------|----------|----------|----------|----------|
| Clinician 1       | 2.63     | 2.82     | 2.81     | 2.90     |
| Clinician 2       | 2.94     | 2.90     | 2.97     | 2.99     |
| Clinician 3       | 2.28     | 2.53     | 2.45     | 2.51     |
| Clinician 4       | 2.94     | 2.75     | 2.80     | 2.83     |

Table 4.15: Summary of Data for Evaluators' (Clinician) Relevance Scores for the SOAP-Note Pools

| <b>Note Pools</b> | <b>Mean</b> | <b>Median</b> | <b>Std. Dev.</b> |
|-------------------|-------------|---------------|------------------|
| K                 | 2.70        | 2.79          | 0.31             |
| E                 | 2.75        | 2.79          | 0.16             |
| U                 | 2.76        | 2.81          | 0.22             |
| T                 | 2.81        | 2.87          | 0.21             |

assessments. Clinician 2 displayed a very narrow range of scores, with all ratings near 3.00. This clinician gave the highest score to pool T (2.99) and the lowest to pool E (2.90), indicating uniformly positive evaluations across the pools. In contrast, Clinician 3 assigned lower scores overall, ranging from 2.28 for pool K to 2.51 for pool T, suggesting a more conservative approach to relevance evaluation. Clinician 4 rated pools K and E similarly (2.94 and 2.75, respectively) but gave slightly higher scores to pools U and T (2.80 and 2.83, respectively). Like Clinician 2, this evaluator tended to provide high ratings across all pools.

Table ?? presents aggregate data on the relevance scores for the SOAP note pools, including the mean, median, and standard deviation. Pool T achieved the highest mean score (2.81), followed by pools U (2.76), E (2.75), and K (2.70). The differences in mean scores across pools were minimal, with a maximum difference of 0.11 between the highest (T) and the lowest (K), indicating all pools were generally perceived as relevant. The median scores further confirm this trend, as pools K and E all had a median of 2.79, while pool T had a slightly higher median of 2.87. The close alignment of mean and median scores suggests symmetrical distributions of relevance scores for all pools.

The standard deviation values provide additional insights into the variability of evaluations. Pool K exhibited the highest variability (0.31), reflecting greater disagreement among clinicians about its relevance. In contrast, pool E had the lowest standard deviation (0.16), indicating high consistency in relevance scores. Pools U and T demonstrated moderate variability, with standard deviations of 0.22 and 0.21, respectively, suggesting some degree of agreement among clinicians.

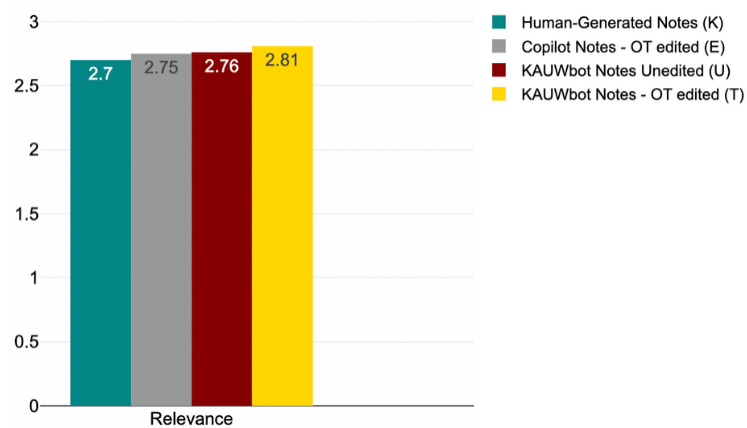


Figure 4.15: Histogram of Mean Relevance Scores for SOAP-Note Pools.

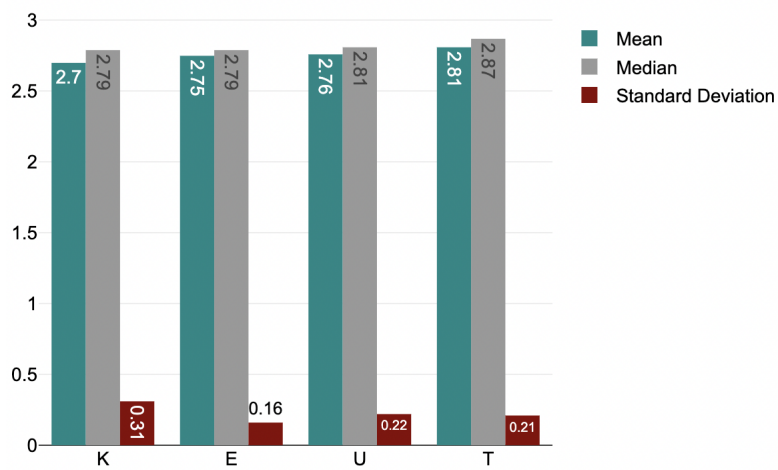


Figure 4.16: Histogram of Mean, Median, and Standard Deviation of Relevance Scores for SOAP-Note Pools.

## 4.7 Organization Performance for the Four SOAP-Note Pools

The descriptive statistics for the evaluators' (clinician) organization scores for the SOAP-note pools, as shown in the tables (Table 4.16 and Table 4.17) below, reveal insights into how the different note pools were perceived in terms of their organization by the clinicians.

Table 4.16: Summary of Mean Values for Evaluators' (Clinician) Organization Scores for the SOAP-Note Pools

| <b>Evaluators</b> | <b>K</b> | <b>E</b> | <b>U</b> | <b>T</b> |
|-------------------|----------|----------|----------|----------|
| Clinician 1       | 1.96     | 2.19     | 2.19     | 2.27     |
| Clinician 2       | 2.36     | 2.73     | 2.54     | 2.74     |
| Clinician 3       | 1.33     | 2.24     | 2.23     | 2.31     |
| Clinician 4       | 2.40     | 2.76     | 2.39     | 2.57     |

Table 4.17: Summary of Data for Evaluators' (Clinician) Organization Scores for the SOAP-Note Pools

| <b>Note Pools</b> | <b>Mean</b> | <b>Median</b> | <b>Std. Dev.</b> |
|-------------------|-------------|---------------|------------------|
| K                 | 2.01        | 2.16          | 0.50             |
| E                 | 2.48        | 2.49          | 0.31             |
| U                 | 2.34        | 2.31          | 0.16             |
| T                 | 2.47        | 2.44          | 0.22             |

Clinician 1 rated pool K the lowest (1.96), followed by similar ratings for pools E and U (2.19) and the highest score for pool T (2.27). The relatively narrow range of scores (1.96 to 2.27) suggests that Clinician 1 perceived slight but consistent differences in the organization among the pools. Clinician 2 gave higher overall scores, with the lowest rating for pool K (2.36) and the highest for pool T (2.74). Pool E (2.73) was rated nearly as highly as T, while pool U (2.54) received a slightly lower score, indicating that this clinician found pools E and T particularly well-organized.

Clinician 3 assigned the lowest scores among all evaluators, with pool K receiving the lowest score (1.33) and pool T the highest (2.31). Pools E (2.24) and U (2.23) received nearly identical scores, suggesting that Clinician 3 perceived limited differentiation between these pools in terms of organization. Clinician 4 provided the most favourable scores, with pool K receiving 2.40 and pool E receiving the highest score (2.76). This clinician rated pool U slightly lower (2.39) and pool T at 2.57, indicating that they viewed pool E as the most organized and pool K as the least.

The mean scores reveal that pool E had the highest average organization score (2.48), followed closely by pool T (2.47) and pool U (2.34). Pool K had the lowest mean score (2.01), indicating it was perceived as the least organized. The median scores closely align with the means, with pool E having the highest median (2.49) and pool K the lowest (2.16). These values suggest that the distribution of scores for each pool is fairly symmetrical.

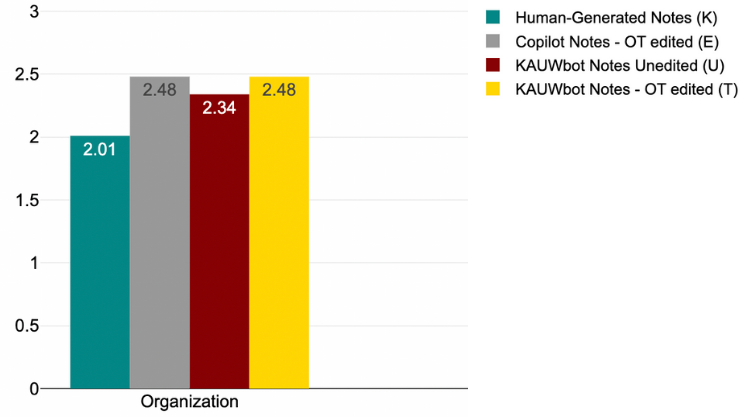


Figure 4.17: Histogram of Mean Relevance Scores for SOAP-Note Pools.

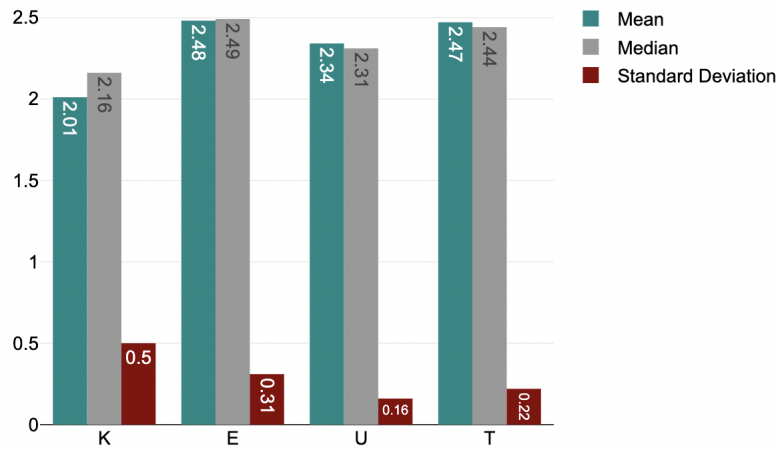


Figure 4.18: Histogram of Mean, Median, and Standard Deviation of Organization Scores for SOAP-Note Pools.

In terms of variability, pool K exhibited the highest standard deviation (0.50), indicating the greatest inconsistency among evaluators regarding its organization. Pool U had the lowest standard deviation (0.16), suggesting strong agreement among evaluators about its organization. Pools E and T demonstrated moderate variability, with standard deviations of 0.31 and 0.22, respectively, reflecting some degree of evaluator consensus.

# Chapter 5

## Discussion

The findings of this research provide significant insights into the role of AI in clinical documentation, particularly in pediatric rehabilitation settings. By systematically comparing human-authored SOAP notes with AI-generated and AI-assisted notes across multiple quality criteria, this research contributes to a growing understanding of how AI can support and potentially enhance clinical documentation practices. This discussion synthesizes the key results, contextualizes them within the existing literature, and explores their implications for clinical practice, training, and future research.

### 5.1 Quality of AI-Generated SOAP Notes

The research revealed that AI-generated SOAP notes, particularly those produced by KAUWbot and subsequently edited by occupational therapists, scored highest across all quality criteria, including clarity, completeness, conciseness, relevance, and organization [5]. This finding shows the potential of AI tools to produce high-quality documentation when augmented by human expertise. The improvement in quality scores for KAUWbot notes after human editing highlights the importance of collaborative models in clinical documentation, where AI serves as a first draft generator and clinicians refine the output to align with clinical standards and personal preferences [5].

Interestingly, unedited AI-generated notes (KAUWbot) outperformed human-authored notes in terms of overall quality. This result suggests that AI models, when trained on domain-specific data, can achieve a level of consistency and accuracy that may surpass human efforts, particularly in fast-paced clinical environments. However, the lower organization scores for unedited notes indicate that while AI excels in generating detailed and relevant content, it may struggle with structuring information in a manner that adheres to established SOAP note conventions.

## 5.2 Human-AI Collaboration

The findings also highlight the transformative potential of human-AI collaboration in clinical documentation. The substantial improvement in KAUWbot-generated notes after editing by OTs reflects the added value of human judgment in refining AI outputs. This aligns with the literature emphasizing the importance of human oversight in mitigating errors, enhancing relevance, and ensuring the contextual appropriateness of AI-generated documentation.

The research shows the necessity of clinician training to optimize the use of AI tools. Clinicians who provided concise and detailed scratch notes generated high-quality SOAP notes more efficiently and easily [39]. This finding suggests that targeted training programs focusing on effective scratch note creation and AI utilization could further enhance the benefits of AI-assisted documentation.

## 5.3 Implications for Clinical Practice

The integration of AI tools like KAUWbot and Copilot into clinical documentation workflows has the potential to address critical challenges in pediatric rehabilitation, including time constraints and clinician burnout [39]. By automating the initial drafting process, these tools can reduce the administrative burden on clinicians, allowing them to allocate more time to direct patient care. However, the variability in time savings observed across clinicians shows the need for personalized training and support to maximize the efficiency gains from AI adoption [39].

From a patient safety perspective, the high clarity and relevance scores of AI-generated notes suggest that these tools can enhance communication among healthcare providers, thereby improving continuity of care. However, the potential for errors or misclassifications in unedited AI outputs [60] highlights the importance of maintaining robust oversight mechanisms to ensure accuracy and reliability.

## 5.4 Limitations and Recommendations for Future Research

Despite the promising results, this research also highlights several challenges associated with AI in clinical documentation. One key limitation is the dependency of AI tools on the quality of input data. The effectiveness of KAUWbot, for example, was contingent on the quality and specificity of scratch notes provided by clinicians. This shows the need for standardized input protocols to ensure consistent performance.

A notable limitation of this research lies in its sample size. Although the dataset included 432 SOAP notes, which is considerable, it may have been insufficient to achieve the statistical power necessary to detect subtle differences in quality among the groups

[5]. With 108 notes per pool, the research may have lacked the sensitivity required to capture meaningful variations in mean scores. Expanding the sample size in future research could bolster the statistical analyses, enabling more robust and definitive conclusions about differences in the quality of SOAP notes across the groups.

Another limitation pertains to the research’s predominant reliance on quantitative methods for assessing the quality of SOAP notes [5]. While the rubric-based evaluations and statistical analyses provided objective and standardized metrics, they may have overlooked insights best captured through qualitative approaches. Incorporating qualitative methodologies in future research could offer deeper perspectives into the subjective experiences and perceptions of occupational therapists and evaluators, particularly regarding the usability, practicality, and contextual relevance of AI-generated SOAP notes.

The use of a 3-point Likert scale for the SOAP note evaluation presents a limitation for this research due to its limited granularity and reduced ability to discriminate between varying levels of quality. Its simplicity could also introduce evaluator bias, as raters might default to the middle option when uncertain, potentially skewing results and reducing assessment reliability. Using a more detailed scale, such as a 5- or 7-point Likert scale, would enhance sensitivity, enable finer distinctions, and improve the validity of the findings.

The findings of this research are specific to pediatric rehabilitation settings, which may limit their applicability to other clinical contexts [5]. Different specialties often have unique documentation requirements and evaluation criteria, making generalizability a challenge. Future research could address this limitation by exploring the use of AI-assisted documentation tools in diverse clinical settings. This would provide a broader understanding of how these tools adapt to varying documentation styles and whether the benefits of human editing are consistent across specialties.

Privacy and data security remain critical concerns, particularly when using AI tools in healthcare settings. While the use of a custom model like KAUWbot mitigates some risks associated with proprietary systems, ongoing efforts are needed to ensure compliance with data protection regulations and safeguard patient confidentiality.

Confirmation bias may arise in the SOAP note generation process as occupational therapy clinicians may unconsciously favour AI-generated notes that align with their existing documentation styles and expectations. This may lead to selective editing or acceptance of certain AI outputs while overlooking errors or inconsistencies. To mitigate this, multiple clinicians could review and refine the AI-generated notes in a multi-step process instead of having the same clinician generate and edit SOAP notes.

# Chapter 6

## Conclusion and Summary

This research has explored the integration of AI in clinical documentation, focusing on the evaluation of AI-generated SOAP notes within pediatric rehabilitation. By comparing human-authored notes with those produced by AI tools—specifically KAUWbot, a custom-trained model, and Copilot, a general-purpose LLM—the research aimed to assess their quality, the extent of required human editing, and their broader implications for clinical practice. The findings show the potential of AI-assisted documentation while highlighting critical areas for improvement and further investigation.

### 6.1 Summary of Key Findings

1. **Quality of AI-Generated SOAP Notes:** AI-generated SOAP notes demonstrated promising levels of clarity, completeness, conciseness, relevance, and organization. The KAUWbot model, fine-tuned specifically for pediatric rehabilitation documentation, generally outperformed the general-purpose Copilot in terms of domain-specific relevance and organization.
2. **Role of Human Editing:** The research revealed that human editing improved the quality of AI-generated SOAP notes. This shows the necessity of clinician involvement to refine and contextualize AI outputs, ensuring they meet clinical and documentation standards.
3. **Statistical Analysis of Quality Metrics:** Through rubric-based evaluations and statistical analyses, variations in quality metrics were observed across different note pools.
4. **Privacy and Ethical Considerations:** The use of AI tools like KAUWbot mitigates some risks associated with proprietary systems, such as data privacy concerns. However, ongoing efforts are essential to ensure compliance with data protection regulations and safeguard patient confidentiality.

# References

- [1] Crissa Nicole Alonso, Genell Lagare, Czyrhen Veanne Chua, Madjoriel Alaan, Selina Jane Escandor, and Abelardo Apollo David. The Family Therapy Engagement Model (FTEM). *Philippine Journal of Allied Health Sciences*, 5(2):77–84, 2022.
- [2] Yasser K Alotaibi and Frank Federico. The Impact of Health Information Technology on Patient Safety. *Saudi Medical Journal*, 38(12):1173, 2017.
- [3] Shuroug A Alowais, Sahar S Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I Alshaya, Sumaya N Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A Badreldin, et al. Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education*, 23(1):689, 2023.
- [4] Mustafa Ahmad Alqaisi and Hassan Izzeddin Sarsak. Occupational Therapy Documentation in Arab Countries: Current Content, Methods, Challenges, and Future Directions. *World Federation of Occupational Therapists Bulletin*, 78(2):90–98, 2022.
- [5] Solomon Amenyo, Maura R Grossman, Daniel G Brown, and Brendan Wylie-Toal. Assessment of AI-Generated Pediatric Rehabilitation SOAP-Note Quality. *arXiv preprint arXiv:2503.15526*, 2025.
- [6] Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, and Yuta Nakamura. Natural Language Processing: From Bedside to Everywhere. *Yearbook of Medical Informatics*, 31(01):243–253, 2022.
- [7] Asha V Asher. Book Review: Best Practices for Documenting Occupational Therapy Services in Schools. *Canadian Journal of Occupational Therapy*, 84:261 – 261, 2017.
- [8] Vijaya Chandra Reddy Avula and Sridhar Amalakanti. Artificial Intelligence in Psychiatry, Present Trends, and Challenges: An Updated Review. *Archives of Mental Health*, 25(1):85–90, 2024.
- [9] Kartikeya Bajpai, Jillian Chown, Gopi Astik, and Kristine Green. Relational Aspects of Regulating Clinical Work: Examining Electronic and In-Person Compliance Mechanisms. *BMJ Leader*, 6:180–185, 2021.
- [10] Hayden P. Baker, Emma P Dwyer, Senthooran Kalidoss, Kelly K. Hynes, Jennifer Moriatis Wolf, and Jason A. Strelzow. ChatGPT’s Ability to Assist with Clinical

Documentation: A Randomized Controlled Trial. *Journal of the American Academy of Orthopaedic Surgeons*, 32:123 – 129, 2023.

- [11] Derek J Baughman, Paul A Botros, and Abdul Waheed. Technology in Medicine: Improving Clinical Documentation. *FP Essentials*, 537:26–38, 2024.
- [12] Arngeir Berge, Frode Guribye, Siri-Linn Schmidt Fotland, Gro Fonnes, Ingrid H Johansen, and Christoph Trattner. Designing for Control in Nurse-AI Collaboration During Emergency Medical Calls. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 1339–1352, 2023.
- [13] Callie M. Berkowitz, Jessica E Ma, Jared Lowe, and Rowena J. Dolor. Assessing Quality in Advance Care Planning Documentation: A Survey of Current Methods. *American Journal of Hospice and Palliative Medicine*, 39:945–950, 2021.
- [14] Anjanava Biswas and Wrick Talukdar. Intelligent Clinical Documentation: Harnessing Generative AI for Patient-Centric Clinical Note Generation. *arXiv preprint arXiv:2405.18346*, 2024.
- [15] Heleen Blijlevens and Juanita F. Murphy. Washing Away SOAP Notes: Refreshing Clinical Documentation. *New Zealand Journal of Occupational Therapy*, 50:3, 2003.
- [16] Antonella Borrelli, Martina Pecoraro, Francesco del Giudice, Leonardo Cristofani, Emanuele Messina, Ailin Dehghanpour, Nicholas Landini, Michela Roberto, Stefano Perotti, Maurizio Muscaritoli, Daniele Santini, Carlo Catalano, and Valeria Panebianco. Standardization of Body Composition Status in Patients with Advanced Urothelial Tumors: The Role of a CT-Based AI-Powered Software for the Assessment of Sarcopenia and Patient Outcome Correlation. *Cancers*, 15(11):2968, 2023.
- [17] Claus Bossen and Kathleen H Pine. Batman and Robin in Healthcare Knowledge Work: Human-AI Collaboration by Clinical Documentation Integrity Specialists. *ACM Transactions on Computer-Human Interaction*, 30(2):1–29, 2023.
- [18] Courtney L Bradley, Karlie Wieder, and Shaina E Schwartz. Shifting from SOAP Notes to Consult Notes for Clinical Documentation by Pharmacy Students. *American Journal of Pharmaceutical Education*, 86(7), 2022.
- [19] Adele Breen-Franklin. The OTA’S Guide to Documentation: Writing SOAP Notes (4th Edition) 2017. *Occupational Therapy In Health Care*, 32:101 – 102, 2018.
- [20] Tracey A. Brereton, Momin M. Malik, Mark Lifson, JAS. Greenwood, Kevin J. Peterson, and Shauna M. Overgaard. The Role of Artificial Intelligence Model Documentation in Translational Science: Scoping Review. *Interactive Journal of Medical Research*, 12(1):e45903, 2023.
- [21] Elizabeth Brey and Kristin Shutts. Children Use Nonverbal Cues to Make Inferences About Social Power. *Child Development*, 86(1):276–286, 2015.

- [22] G. Ted Brown, Sylvia A. Rodger, Anita B O Brown, and Carsten Roever. A Profile of Canadian Pediatric Occupational Therapy Practice. *Occupational Therapy In Health Care*, 21(4):39–69, 2007.
- [23] Javier Cámara, Javier Troya, Lola Burgueño, and Antonio Vallecillo. On the assessment of generative AI in modeling tasks: an experience report with ChatGPT and UML. *Software and Systems Modeling*, 22(3):781–793, 2023.
- [24] Alejandra Caqueo-Urizar, Alfonso Urzua, Ester Villalonga-Olives, Diego Atencio-Quevedo, Matías Irrázaval, Jerome Flores, and Cristian Ramirez. Children’s Mental Health: Discrepancy Between Child Self-Reporting and Parental Reporting. *Behavioral Sciences*, 12(10):401, 2022.
- [25] Jessica P Cerdeña, Marie V Plaisime, and Jennifer Tsai. From Race-Based to Race-Conscious Medicine: How Anti-Racist Uprisings Call Us to Act. *The Lancet*, 396(10257):1125–1128, 2020.
- [26] Stephen J. Chapman. Review of Discovering Statistics Using IBM SPSS Statistics, 4th Edition. *Journal of Political Science Education*, 14:145–147, 2018.
- [27] Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H Kann, Fallon E Chipidza, Jonathan Leeman, Hugo JWL Aerts, Timothy Miller, et al. The Impact of Responding to Patient Messages with Large Language Model Assistance. *arXiv preprint arXiv:2310.17703*, 2023.
- [28] Shih-Ying Chien, Alice Mk Wong, Ching-Yi Wu, and Sara L Beckman. Interactive Electronic Pegboard for Enhancing Manual Dexterity and Cognitive Abilities: Instrument Usability Study. *JMIR Human Factors*, 11:e56357, 2024.
- [29] Suhail Chughtai. AI-empowered Emergency Department Cutting Edge Digital Medicine. *Journal of Pakistan Society of Internal Medicine*, 5(4):758–761, 2024.
- [30] Enrico Coiera, Baki Kocaballi, John Halamka, and Liliana Laranjo. The Digital Scribe. *NPJ Digital Medicine*, 1(1):58, 2018.
- [31] Brian Creaner and Clodagh Nolan. How Do We Know We Do What It Says on the Tin? Ensuring Adherence to a Person-Centered Approach Within Occupational Therapy Mental Health Services in Higher Education. *European Journal for Person Centered Healthcare*, 4:271–280, 2016.
- [32] Astrid Crom, Davy Paap, Amarins J. Wijma, Pieter U Dijkstra, and Grieteke Pool. Between the Lines: A Qualitative Phenomenological Analysis of the Therapeutic Alliance in Pediatric Physical Therapy. *Physical & Occupational Therapy In Pediatrics*, 40(1):1–14, 2020.
- [33] Aine Cryts. The Pros and Cons of Using Templates in the EHR. <https://www.physicianspractice.com/view/pros-and-cons-using-templates-ehr>, 2016. [Online; accessed 10-February-2025].

- [34] Nadia Daneshvar, Deepti Pandita, Shari M. Erickson, Lois Snyder Sulmasy, and Matthew DeCamp. Artificial Intelligence in the Provision of Health Care: An American College of Physicians Policy Position Paper. *Annals of Internal Medicine*, 177(7):964–967, 2024.
- [35] Oswalcir Almeida de Azevedo and Diná Cruz. Quality Indicators of the Nursing Process Documentation in Clinical Practice. *Revista Brasileira de Enfermagem*, 74(3):e20201355, 2021.
- [36] Alex J. DeGrave, Zhuo Cai, Joseph D. Janizek, Roxana Daneshjou, and Su-In Lee. Dissection of Medical AI Reasoning Processes via Physician and Generative-AI Collaboration. *medRxiv*, 2023.
- [37] Paul Denny, Viraj Kumar, and Nasser Giacaman. Conversing with Copilot: Exploring Prompt Engineering for Solving CS1 Problems Using Natural Language. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 1136–1142, 2023.
- [38] Giovanni Di Napoli and Linda S Lee. The Brave New World of Artificial Intelligence: Dawn of a New Era. *iGIE*, 2(1):62–69, 2023.
- [39] Rachel DiMaio. Creation of a Custom Language Model for Pediatric Occupational Therapy Documentation. Master’s Thesis, University of Waterloo, 2024.
- [40] Chelsea Doktorchik, Mingshan Lu, Hude Quan, Cathy Ringham, and Cathy Eastwood. A Qualitative Evaluation of Clinically Coded Data Quality from Health Information Manager Perspectives. *Health Information Management Journal*, 49(1):19–27, 2020.
- [41] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. Opinion Paper: “So What if ChatGPT Wrote It?” Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. *International Journal of Information Management*, 71:102642, 2023.
- [42] Kees Ebben, Melle Sieswerda, Ernest J. T. Luiten, Joan B. Heijns, Carmen C. van der Pol, Maud Bessems, Aafke H Honkoop, Mathijs P. Hendriks, Janneke Verloop, and Xander A. Verbeek. Impact on Quality of Documentation and Workload of the Introduction of a National Information Standard for Tumor Board Reporting. *JCO Clinical Cancer Informatics*, 4:346–356, 2020.
- [43] Christof Ebert and Panos Louridas. Generative AI for Software Practitioners. *IEEE Software*, 40(4):30–38, 2023.
- [44] Joanne Phillips Estes and Doris Pierce. Pediatric Therapists’ Perspectives on Occupation-Based Practice. *Scandinavian Journal of Occupational Therapy*, 19(1):17–25, 2012.

- [45] Jonah Feldman, Adam Goodman, Katherine Hochman, Eesha Chakravartty, Jonathan Austrian, Eduardo Iturrate, Brian Bosworth, Archana Saxena, Marwa Moussa, Dina Chenouda, et al. Novel Note Templates to Enhance Signal and Reduce Noise in Medical Documentation: Prospective Improvement Study. *JMIR Formative Research*, 7:e41223, 2023.
- [46] Teresa Foy, Ginger Perritt, Deepa Thimmaiah, Lauren Heisler, Jennifer Lookingbill Offutt, Kara Cantoni, Ching-Hui Hseih, Julie Gassaway, Rebecca Ozelie, and Deborah Backus. Occupational Therapy Treatment Time During Inpatient Spinal Cord Injury Rehabilitation. *The Journal of Spinal Cord Medicine*, 34(2):162–175, 2011.
- [47] Chris Giordano, Meghan Brennan, Basma Mohamed, Parisa Rashidi, François Modave, and Patrick Tighe. Accessing Artificial Intelligence for Clinical Decision-Making. *Frontiers in Digital Health*, 3:645232, 2021.
- [48] Peter Glen, Naomi Earl, Felix Gooding, Emily Lucas, Nicole Sangha, and Steve Ramcharitar. Simple Interventions Can Greatly Improve Clinical Documentation: A Quality Improvement Project of Record Keeping on the Surgical Wards at a District General Hospital. *BMJ Open Quality*, 4(1):u208191–w3260, 2015.
- [49] Paulette Guitard. Book Review: Documentation Manual for Occupational Therapy: Writing SOAP Notes. *Canadian Journal of Occupational Therapy*, 80:180–180, 2013.
- [50] Antje Hammer, Anke Wagner, Monika Annemarie Rieger, and Tanja Manser. Assessing the Quality of Medication Documentation: Development and Feasibility of the MediDocQ Instrument for Retrospective Chart Review in the Hospital Setting. *BMJ Open*, 9(11):e034609, 2019.
- [51] Patricia Hay, Kathy Wilton, Jennifer Barker, Julie Mortley, and Megan Cumerlato. The Importance of Clinical Documentation Improvement for Australian Hospitals. *Health Information Management Journal*, 49:69–73, 2019.
- [52] Jim Hinojosa, Jill Anderson, and Carrie Strauch. Pediatric Occupational Therapy in the Home. *The American Journal of Occupational Therapy*, 42(1):17–22, 1988.
- [53] Y-X Ho, CS Gadd, KL Kohorst, and ST Rosenbloom. A Qualitative Analysis Evaluating the Purposes and Practices of Clinical Documentation. *Applied Clinical Informatics*, 5(01):153–168, 2014.
- [54] Rebecca Miriam Jedwab, Michael Franco, Denise Owen, Anna Ingram, Bernice Redley, and Naomi Dobroff. Improving the Quality of Electronic Medical Record Documentation: Development of a Compliance and Quality Program. *Applied Clinical Informatics*, 13(04):836–844, 2022.
- [55] Gregor Jošt, Viktor Taneski, and Sašo Karakatič. The Impact of Large Language Models on Programming Education and Student Learning Outcomes. *Applied Sciences*, 14(10):4115, 2024.

- [56] Hyoje Jung, Yunha Kim, Heejung Choi, Hyeram Seo, Minkyung Kim, Jiye Han, Gaeun Kee, Seohyun Park, Soyoung Ko, Byeolhee Kim, Suyeon Kim, Tae Joon Jun, and Young-Hak Kim. Enhancing Clinical Efficiency through LLM: Discharge Note Generation for Cardiac Patients. *arXiv preprint arXiv:2404.05144*, 2024.
- [57] Andreas Jungherr. Using ChatGPT and Other Large Language Model (LLM) Applications for Academic Paper Assignments. *SocArXiv*, 2023.
- [58] Joyce Kamanzi, Abraham Megentta, Wilson Nsabiyumva, Augustin Sendegeya, and Rex Wong. Improving Clinical Documentation Through Monthly Audits in Butare Teaching Hospital, Rwanda. *Journal of Service Science and Management*, 8(6):860–867, 2015.
- [59] Leo Kelion. Babylon Health Admits GP App Suffered a Data Breach. <https://www.bbc.com/news/technology-52986629>, 2020. [Online; accessed 05-February-2025].
- [60] Annessa Kernberg, Jeffrey A Gold, and Vishnu Mohan. Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study. *Journal of Medical Internet Research*, 26:e54419, 2024.
- [61] Arjun Kharpal. Google DeepMind Patient Data Deal with UK Health Service Illegal, Watchdog Says. <https://www.cnbc.com/2017/07/03/google-deepmind-nhs-deal-health-data-illegal-ico-says.html>, 2017. [Online; accessed 05-February-2025].
- [62] KidsAbility. KidsAbility Centre for Child Development. <https://www.kidsability.ca/about-us/>, 2024. [Online; accessed 12-February-2025].
- [63] Anne Kiraly-Alvarez. Assessing Volition in Pediatrics: Using the Volitional Questionnaire and the Pediatric Volitional Questionnaire. *The Open Journal of Occupational Therapy*, 3(3):7, 2015.
- [64] Bonnie Klassen, Lili Liu, and Sharon Warren. Pain Management Best Practice with Older Adults: Effects of Training on Staff Knowledge, Attitudes, and Patient Outcomes. *Physical & Occupational Therapy In Geriatrics*, 27:173 – 196, 2009.
- [65] Ahmet Baki Kocaballi, Kiran Ijaz, Liliana Laranjo, Juan C. Quiroz, Dana Rezazadegan, Huong Ly Tong, Simon Willcock, Shlomo Berkovsky, and Enrico W. Coiera. Envisioning an Artificial Intelligence Documentation Assistant for Future Primary Care Consultations: A Co-Design Study with General Practitioners. *Journal of the American Medical Informatics Association*, 27(11):1695–1704, 2020.
- [66] Laxminarayana Korada. GitHub Copilot: The Disrupting AI Companion Transforming the Developer Role and Application Lifecycle Management. *Journal of Artificial Intelligence & Cloud Computing*, 348:2–4, 2024.

- [67] Heiko Koziolk, Sten Gruener, and Virendra Ashiwal. ChatGPT for PLC/DCS Control Logic Generation. In *2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–8. IEEE, 2023.
- [68] Hanne Kaae Kristensen, Charlotte Ytterberg, Dorrie Jones, and Hans Lund. Research-Based Evidence in Stroke Rehabilitation: An Investigation of its Implementation by Physiotherapists and Occupational Therapists. *Disability and Rehabilitation*, 38:2564 – 2574, 2016.
- [69] Kyle Lam. ChatGPT for Low-and Middle-Income Countries: A Greek Gift? *The Lancet Regional Health–Western Pacific*, 41, 2023.
- [70] Chanseo Lee, Sonu Kumar, Kimon A. Vogt, and Sam Meraj. Improving Clinical Documentation with AI: A Comparative Study of Sporo AI Scribe and GPT-4o mini. *arXiv preprint arXiv:2410.15528*, 2024.
- [71] Sangil Lee, Nicholas M Mohr, W Nicholas Street, and Prakash Nadkarni. Machine Learning in Relation to Emergency Medicine Clinical and Operational Scenarios: An Overview. *Western Journal of Emergency Medicine*, 20(2):219, 2019.
- [72] Siaw-Teng Liaw, Harshana Liyanage, Craig Kuziemy, Amanda L Terry, Richard Schreiber, Jitendra Jonnagaddala, and Simon de Lusignan. Nurses Partnering with Technology to Predict the Patients’ Risk of Falling Using Artificial Intelligence. *Yearbook of Medical Informatics*, 29(01):051–057, 2020.
- [73] Chunyu Liu, Yongpei Ma, Kavitha Kothur, Armin Nikpour, and Omid Kavehei. BioSignal Copilot: Leveraging the Power of LLMs in Drafting Reports for Biomedical Signals. *MedRxiv*, 2023.
- [74] Tsai-Ling Liu, Timothy C. Hetherington, Casey Stephens, Andrew McWilliams, Ajay Dharod, Tracey Carroll, and Jeffrey A. Cleveland. AI-Powered Clinical Documentation and Clinicians’ Electronic Health Record Experience. *JAMA Network Open*, 7(9):e2432460–e2432460, 2024.
- [75] Ihab Mansoor, Javier García Ortiz, and Matthew Rector. Generative AI in Clinical Research: Regulatory Submissions, Clinical Data Management, and Beyond. *American Medical Writers Association Journal*, 39(1), 2024.
- [76] Anastasia Raquel Martinez and Karen Jacobs. Establishing a Pediatric Occupational Therapy Clinic in the Public-Sector in Trinidad: The Quest for Clinical and Professional Expansion. *A Journal of Prevention, Assessment & Rehabilitation*, 66(3):669–674, 2020.
- [77] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

- [78] Muhammad Zahaib Nabeel. Big Data Analytics-Driven Project Management Strategies. *Journal of Science & Technology*, 5(1):117–163, 2024.
- [79] Omer Nasim, Abdullah Durrani, Boulos Eskander, Charalampos Pantelias, and Kieran Gallagher. A Clinical Audit of Orthopaedic Operation Note Documentation and Digitalization of the Operative Note Template: A Quality Improvement Project. *Cureus*, 14(12), 2022.
- [80] Suryakiran Navath. Transforming Healthcare: the Impact and Future of Artificial Intelligence in Healthcare. *Journal of Artificial Intelligence and Machine Learning*, 1(1):16–21, 2023.
- [81] Khalid Nawab. Artificial Intelligence Scribe: A New Era in Medical Documentation. *Artificial Intelligence in Health*, 1(4):12–15, 2024.
- [82] Josh Nguyen and Christopher A Pepping. The Application of ChatGPT in Healthcare Progress Notes: A Commentary from a Clinical and Research Perspective. *Clinical and Translational Medicine*, 13(7):e1324, 2023.
- [83] Iona Novak and Ingrid Honan. Effectiveness of Paediatric Occupational Therapy for Children with Disabilities: A Systematic Review. *Australian Occupational Therapy Journal*, 66(3):258–273, 2019.
- [84] Laurie Lovett Novak, Regina G Russell, Kim V. Garvey, Mehool Patel, Kelly Jean Thomas Craig, Jane Snowdon, and Bonnie Miller. Clinical Use of Artificial Intelligence Requires AI-Capable Organizations. *JAMIA Open*, 6(2):ooad028, 2023.
- [85] Robert A Oliver, Jill M Cancio, Christopher A Rábago, and Kathleen E Yancosek. Impact of Fire Arms Training in a Virtual Reality Environment on Occupational Performance (Marksmanship) in a Polytrauma Population. *Military Medicine*, 184(11-12):832–838, 2019.
- [86] Patricia F Pearce, Laurie Anne Ferguson, Gwen S George, and Cynthia A Langford. The Essential SOAP Note in an EHR Age. *The Nurse Practitioner*, 41(2):29–36, 2016.
- [87] Diamond E Rashad, Susan A. Zapf, and Sandra E Groger. The Opportunity for Occupational Therapy in Pediatric Disaster Recovery. *The American Journal of Occupational Therapy*, 76(4), 2022.
- [88] Stacey Reynolds, Tara J Glennon, Karla Ausderau, Roxanna M Bendixen, Heather Miller Kuhanek, Beth Pfeiffer, Renee Watling, Kimberly Wilkinson, and Stefanie C Bodison. Using a Multifaceted Approach to Working with Children Who Have Differences in Sensory Processing and Integration. *The American Journal of Occupational Therapy*, 71(2):7102360010p1–7102360010p10, 2017.
- [89] Fatima Rodriguez, David Scheinker, and Robert A Harrington. Promise and Perils of Big Data and Artificial Intelligence in Clinical Medicine and Biomedical Research. *Circulation Research*, 123(12):1282–1284, 2018.

- [90] Jacquelyn Roth, Sukruthi Thunga, and Jane Yoo. Assessing the Landscape of AI-Powered Patient Documentation in Dermatology. *Journal of Drugs in Dermatology*, 24(1):64–69, 2025.
- [91] Ruthann Russo, Stephen P. Fitzgerald, JD Eveland, Barry D. Fuchs, and D.Patrick Redmon. Improving Physician Clinical Documentation Quality: Evaluating Two Self-Efficacy-Based Training Programs. *Health Care Management Review*, 38(1):29–39, 2013.
- [92] Jörn M Schattenberg, Naga Chalasani, and Naim Alkhouri. Artificial Intelligence Applications in Hepatology. *Clinical Gastroenterology and Hepatology*, 21(8):2015–2025, 2023.
- [93] Heidi Schwellnus, Yukari Seko, Gillian King, Patricia Baldwin, and Michelle Servais. Solution-Focused Coaching in Pediatric Rehabilitation: Perceived Therapist Impact. *Physical & Occupational Therapy in Pediatrics*, 40(3):263–278, 2020.
- [94] Kathrin Seibert, Dominik Domhoff, Dominik Bruch, Matthias Schulte-Althoff, Daniel Fürstenau, Felix Biessmann, and Karin Wolf-Ostermann. Application Scenarios for Artificial Intelligence in Nursing Care: Rapid Review. *Journal of Medical Internet Research*, 23(11), 2021.
- [95] Jan F. Senge, Matthew McMurray, Fabian Haupt, Philipe Sebastian Breiding, Claus Beisbart, Keivan Daneshvar, Alois Komarek, Gerd Nöldge, Frank Mosler, and Wolfram Andreas Bosbach. ChatGPT May Free Time Needed by the Interventional Radiologist for Administration/Documentation: A Study on the RSNA PICC Line Reporting Template. *Swiss Journal of Radiology and Nuclear Medicine*, 7(2), 2024.
- [96] Emre Sezgin, Joseph Sirrianni, and Kelly Kranz. Development and Evaluation of a Digital Scribe: Conversation Summarization Pipeline for Emergency Department Counseling Sessions towards Reducing Documentation Burden. *medRxiv*, 2023.
- [97] Joel Shoolin, L Ozeran, Claus Hamann, and W Bria Ii. Association of Medical Directors of Information Systems Consensus on Inpatient Electronic Health Record Documentation. *Applied Clinical Informatics*, 4(02):293–303, 2013.
- [98] James E Siegler, Neha N Patel, and C Jessica Dine. Prioritizing Paperwork Over Patient Care: Why Can’t We Do Both? *Journal of Graduate Medical Education*, 7(1):16–18, 2015.
- [99] Nina Singh, Katharine Lawrence, Safiya Richardson, and Devin M Mann. Centering Health Equity in Large Language Model Deployment. *PLOS Digital Health*, 2(10):e0000367, 2023.
- [100] Alysha Skuthan and Wendy B Stav. Profile of Assessments in Pediatric Occupational Therapy Practice. *The American Journal of Occupational Therapy*, 77(5):7705205010, 2023.

- [101] Daniel Stephan, Annika Sophie Bertsch, Matthias Burwinkel, Shankeeth Vinayalingam, Bilal Al-Nawas, Peer Wolfgang Kämmerer, and Daniel G. E. Thiem. AI in Dental Radiology—Improving the Efficiency of Reporting With ChatGPT: Comparative Study. *Journal of Medical Internet Research*, 26:e60684, 2024.
- [102] Peter D Stetson, Suzanne Bakken, Jesse O Wrenn, and Eugenia L Siegler. Assessing Electronic Note Quality Using the Physician Documentation Quality Instrument (PDQI-9). *Applied Clinical Informatics*, 3(02):164–174, 2012.
- [103] Peter D Stetson, Frances P Morrison, Suzanne Bakken, Stephen B Johnson, and eNote Research Team. Preliminary Development of the Physician Documentation Quality Instrument. *Journal of the American Medical Informatics Association*, 15(4):534–541, 2008.
- [104] Deephak Swaminath, Jennifer Hefner, Leigh Ann Jenkins, Jose A Suarez, Gary Meyerrose, and Timothy R Huerta. Improving Clinical Documentation in an Academic Setting: A Multidisciplinary Team Approach Focused on the Physician’s Perspective. *Journal of Healthcare Management*, 63(5):e88–e98, 2018.
- [105] Amelia Tan, Jodie A Copley, and Jennifer M. Fleming. Clinical Utility of a Decision-Making Aid for Upper Limb Neurorehabilitation: Applying the Hypertonicity Intervention Planning Model Across Cultures. *Brain Injury*, 37(7):572–580, 2023.
- [106] Chee Wei Tan, Shangxin Guo, Man Fai Wong, and Ching Nam Hang. Copilot for Xcode: Exploring AI-Assisted Programming by Prompting Cloud-based Large Language Models. *arXiv preprint arXiv:2307.14349*, 2023.
- [107] Oliver Y Tang, Rohaid Ali, Ian D Connolly, Jared S Fridley, Patricia L Zadnik Sullivan, Deus Cielo, Adetokunbo A Oyelese, Curtis E Doberstein, Albert E Telfeian, Ziya L Gokaslan, et al. The Urgency of Neurosurgical Leadership in the Era of Artificial Intelligence. *Neurosurgery*, 93(3), 2023.
- [108] Murat Tepe and Emre Emekli. Assessing the Responses of Large Language Models (ChatGPT-4, Gemini, and Microsoft Copilot) to Frequently Asked Questions in Breast Imaging: A Study on Readability and Accuracy. *Cureus*, 16(5), 2024.
- [109] Sarah Traeger. Using a Dental Clinical Notes Template: Pros and Cons. <https://www.dentalcashflowsolutions.com/blog/using-a-dental-clinical-notes-template-pros-and-cons>, 2023. [Online; accessed 10-February-2025].
- [110] Brian D Tran, Kareem Latif, Tera L Reynolds, Jihyun Park, Jennifer Elston Lafata, Ming Tai-Seale, and Kai Zheng. “Mm-hm,” “Uh-uh”: Are Non-Lexical Conversational Sounds Deal Breakers for the Ambient Clinical Documentation Technology? *Journal of the American Medical Informatics Association*, 30(4):703–711, 2023.
- [111] Derick T Wade. What is Rehabilitation? An Empirical Investigation Leading to an Evidence-based Description. *Clinical Rehabilitation*, 34(5):571–583, 2020.

- [112] I. V. Wagner. Quality Assurance and Quality Enhancement by Use of Digital Clinical Documentation. *International Journal of Computerized Dentistry*, 9(3):195–204, 2006.
- [113] Yophtahe B. Woldegerima and Semira D. Kemal. Clinical Audit on the Practice of Documentation at Preanesthetic Evaluation in a Specialized University Hospital. *Anesthesia, Essays and Researches*, 12:819–824, 2018.
- [114] Polly Yeung. SPSS Survival Manual. A Step by Step Guide to Data Analysis Using IBM SPSS (5th ed) [Book Review]. *Aotearoa New Zealand Social Work*, 26:92, 2014.
- [115] Vahid Zolfaghari, Nenad Petrovic, Fengjunjie Pan, Krzysztof Lebioda, and Alois Knoll. Adopting RAG for LLM-Aided Future Vehicle Design. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 437–442. IEEE, 2024.