

Master's Programme in Biomedical Engineering (BME)

Enhancing Pharmacy IT Support with a Conversational AI Agent

Pharmadata Case Study: Improving First-Line Customer Service Operations

Elisa Ståhlberg

Master's thesis
2025

Copyright ©2025 Elisa Ståhlberg

Author	Elisa Ståhlberg		
Title of thesis	Enhancing Pharmacy IT Support with a Conversational AI Agent		
Programme	Master's Programme in Life Science Technologies		
Major	Biomedical Engineering		
Thesis supervisor	Associate Professor Matias Palva		
Thesis advisor(s)	Tiina Kurimo, MBA		
Collaborative partner	Pharmadata Oy		
Date	Number of pages	Language	
19.12.2025	92 + 7	English	

Abstract

Digital transformation in the pharmacy sector has increased the need for scalable and efficient customer support solutions. Advances in Conversational AI (CAI) and large language models (LLMs) offer new opportunities to revolutionise customer service operations by strengthening efficiency and scalability. While the potential of CAI technology is widely recognised its practical use in regulated healthcare information technology (IT) remains unstudied. This thesis examines how LLM-based CAI can support and partly automate first-line customer support in pharmacy IT. The study was conducted with Pharmadata Oy, a Finnish provider of pharmacy IT solutions. The empirical investigation explores CAI's suitability for first-line pharmacy IT support by combining structured analysis of real support tickets with targeted LLM performance evaluations. The analysis identifies which task types are suitable for safe CAI automation, which require human judgement and the key organisational, technical, and data factors that shape successful adoption. The primary outcome of this research is an empirically grounded assessment of CAIs performance in first-line pharmacy IT support, with a focus on accuracy and escalation behaviour. The findings identify tasks suited to reliable CAI support as well as situations where human involvement remains necessary. Together these results offer a foundation for future development and support the safe and reliable integration of CAI into pharmacy IT workflows.

The results indicated that CAI performs best in predictable and well-documented workflows, such as retrieving information and guiding users through routine tasks. The performance evaluation also showed that several high-frequency queries were successfully automated even in categories that were otherwise less suitable for CAI, highlighting the need for task-level rather than category-level assessment. Tasks requiring dynamic system-state information, sensitive data, or complex professional reasoning remained unsuitable for automation. The analysis further suggests that CAI's strongest contribution lies in streamlining routine support activities, which allows human specialists to focus on tasks requiring deeper expertise. The study also identifies several organisational and technical factors that shape successful CAI use, including clear workflow documentation and reliable access to support materials.

Keywords Conversational AI (CAI), Pharmacy IT, Digital Transformation, Large Language Models

Tekijä Elisa Ståhlberg

Työn nimi Keskustelevan tekoälyn hyödyntäminen apteekkien IT-tuen tehostamisessa

Koulutusohjelma Master's Programme in Life Science Technologies

Pääaine Biomedical Engineering

Vastuuopettaja/valvoja Prof. TkT Matias Palva

Työn ohjaaja(t) Tiina Kurimo, MBA

Yhteistyötaho Pharmadata Oy

Päivämäärä 19.12.2025 **Sivumäärä** 92 + 7

Kieli Englanti

Tiivistelmä

Apteekkilan digitaalinen transformaatio on lisännyt tarvetta tehokkaille asiakastuen ratkaisuille. Keskustelevan tekoälyn (CAI) ja suurten kielimallien (LLM) kehitys tarjoaa uusia mahdollisuuksia tehostaa asiakaspalvelua parantamalla prosessien tehokkuutta ja skaalautuvuutta. Vaikka CAI-tekniikan potentiaali on laajasti tunnustettu, sen käytännön soveltamista säännellyssä terveydenhuollon ympäristössä on tutkittu vain rajallisesti. Tämä diplomityö tarkastelee, miten LLM-pohjainen CAI voi tukea ja osittain automatisoida apteekkien IT-järjestelmien asiakastukea. Tutkimus toteutettiin yhteistyössä suomalaisen apteekkilan IT-palveluntarjoaja Pharmadata Oy:n kanssa. Empiirinen tutkimus arvioi CAI:n soveltuvuutta apteekkilan IT-tukeen yhdistämällä historiallisten tukipyyntöjen analyysin LLM-suoritustestiin, jossa kielimallin vastaukset pisteytettiin todellisiin tukipyyntöihin perustuvissa tehtävissä. Analyysi osoittaa, mitkä tehtävät soveltuvat turvallisesti CAI-automaatiolle ja milloin ihmisen harkinta on edelleen välttämätöntä. Lisäksi tutkimus tunnistaa keskeiset tekniset ja dataperusteiset edellytykset, jotka vaikuttavat CAI:n onnistuneeseen käyttöönottoon.

Tutkimuksen keskeisin tulos on empiirinen arvio CAI:n toimivuudesta apteekkien IT-tuen ensilinjan tehtävissä, erityisesti tarkkuuden ja eskaloitukäyttäytymisen näkökulmasta. Tulokset tunnistavat tehtävätyypit, joissa CAI toimii luotettavasti, sekä tilanteet, joissa asiantuntijan osallistuminen on edelleen tarpeen. Tulokset osoittavat, että CAI suoriutuu parhaiten ennakoitavissa ja hyvin dokumentoiduissa tehtävissä, kuten tiedonhaussa ja käyttäjien ohjaamisessa rutiininomaisten työvaiheiden läpi. LLM-suoritustestit osoittivat, että joitain paljon esiintyviä kysymyksiä voitiin automatisoida luotettavasti myös kategorioissa, jotka eivät kokonaisuutena olleet otollisia automaatiolle. Tämä korostaa tehtäväkohtaisen tarkastelun merkitystä kategoriasidonnaisten arviointien sijaan. Tehtävät, jotka vaativat ajantasaista järjestelmätilaa tai arkaluontoisten tietojen käsittelyä, eivät soveltuneet automaatointiin yhtä hyvin kuin rutiinitehtävät. Lisäksi analyysi osoittaa, että CAI:n merkittävin lisäarvo syntyy tukipyyntöjen virtaviivaistamisesta, jolloin asiantuntijat voivat keskittyä vaativampiin tukitehtäviin. Useat tekniset tekijät, kuten selkeä työkulkujen dokumentointi ja luotettava pääsy tukimateriaaleihin vaikuttavat keskeisesti CAI:n onnistuneeseen hyödyntämiseen.

Avainsanat keskusteleva tekoäly, apteekkien IT-järjestelmät, digitaalinen transformaatio, suuret kielimallit

Table of contents

Preface.....	7
Abbreviations.....	8
1 Introduction	9
1.1 Background	9
1.1.1 Limited CAI adoption in pharmacy IT support.....	10
1.2 Topic and scope	11
1.2.1 Research questions	12
1.3 Contribution to existing literature	13
1.4 Structure of the thesis.....	15
1.5 Research approach and methods	16
2 Literature review	18
2.1 Digital transformation in the pharmacy sector.....	18
2.2 CAI systems.....	19
2.2.1 LLMs as the core of modern CAI	20
2.2.2 CAI system architecture and components	21
2.2.3 Evolving agent interactions.....	23
2.2.4 Automation and decision-support capabilities.....	24
2.2.5 Technical Challenges and Performance Factors.....	25
2.3 Human-CAI interaction.....	28
2.3.1 Prompting and user experience patterns.....	28
2.3.2 Training, onboarding and reducing user burden	30
2.4 Challenges and ethical considerations in CAI support	30
3 Case study – Pharmadata Oy.....	33
3.1 Introduction to case study	33
3.1.1 Analysis design.....	34
3.2 Overview of Pharmadata’s customer support environment	36
3.2.1 Support demand by system platform.....	37
3.2.2 Support demand by functional area.....	38
3.2.3 Overview of structural patterns in support demand	40
3.3 Determining escalation boundaries between CAI and human support agents.....	41

3.4	Test implementation: LLM response evaluation	44
3.4.1	Evaluation framework.....	44
3.4.2	Methods	45
3.4.3	Test data and case selection	47
3.4.4	Data recording and analysis.....	49
4	Results	50
4.1	CAI accuracy varies across categories	52
4.2	Escalation behaviour and boundary recognition	56
4.2.1	Performance in escalation logic	56
4.2.2	Interaction between escalation, accuracy and helpfulness ...	58
4.3	Suitability of CAI for service desk support in pharmacy IT.....	59
4.3.1	Language performance as a precondition for suitability.....	60
4.3.2	Functional category-level performance differences	62
4.3.3	Tasks suitable for CAI automation.....	65
4.3.4	Regulatory and operational constraints shaping suitability ..	67
4.3.5	Patterns relevant to reliability.....	68
4.4	System-level constraints affecting CAI performance.....	69
5	Conclusion and discussion.....	71
5.1	Discussion of key findings	71
5.2	Recommendations	77
5.2.1	Establishing safe and effective CAI-supported workflows	78
5.2.2	Strengthening data and technical implementation	79
5.3	Critical reflections on CAI in pharmacy IT support.....	81
5.3.1	Limitations of CAI	82
5.3.2	Ethical considerations	84
5.3.3	Directions for future research	85
	References.....	88
	Appendix A. Evaluation Framework for Copilot Performance.....	93
	Appendix B. Copilot CAI performance evaluation data.....	94
	Appendix C. Distribution of test queries requiring escalation versus automation.....	96
	Appendix D. Evaluation of escalation decisions by category and query	97

Preface

As I bring my studies to a close with this thesis, I find myself reflecting on a journey that began under unusual circumstances. I started my studies during the pandemic, a time marked by uncertainty and distance, which made the following years feel especially meaningful. Since then, my path has been anything but straightforward. I changed majors, spent time working abroad, and returned to university with a clearer sense of what I wanted to pursue. That decision led me to a field that truly captured my interest and working with topics connected to the pharmaceutical world as well as AI has been both challenging and rewarding.

I want to thank Pharmadata for giving me the opportunity to work on this subject and for supporting the research behind this thesis. I appreciate everyone at Pharmadata who contributed insights that helped me understand the practical world behind the technology. I am especially grateful to my thesis advisor Tiina Kurimo and supervisor Matias Palva for their guidance and for the time they invested throughout this work. I would also like to acknowledge the Teknologiateollisuuden 100-vuotissäätiö, whose financial support has contributed to the completion of this thesis.

Most of all, I want to thank my friends, partner, and family. Your encouragement and belief in me made it possible to continue even when the work felt overwhelming. Thank you to everyone who made these years unforgettable. It has been quite the ride, and I cannot wait to see what the future has in store.

Helsinki, December 19th, 2025

Elisa Ståhlberg

Abbreviations

AI	Artificial Intelligence
AI TRiSM	AI Trust, Risk and Security Management
BPM	Business Process Management
CAI	Conversational Artificial Intelligence
EA	Enterprise Architecture
GenAI	Generative Artificial Intelligence
GDPR	General Data Protection Regulation
IT	Information Technology
LLM	Large Language Model
MAS	Multi-Agent System
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
PoC	Proof of Concept
RAG	Retrieval-Augmented Generation
SoP	Standard operating Procedure

1 Introduction

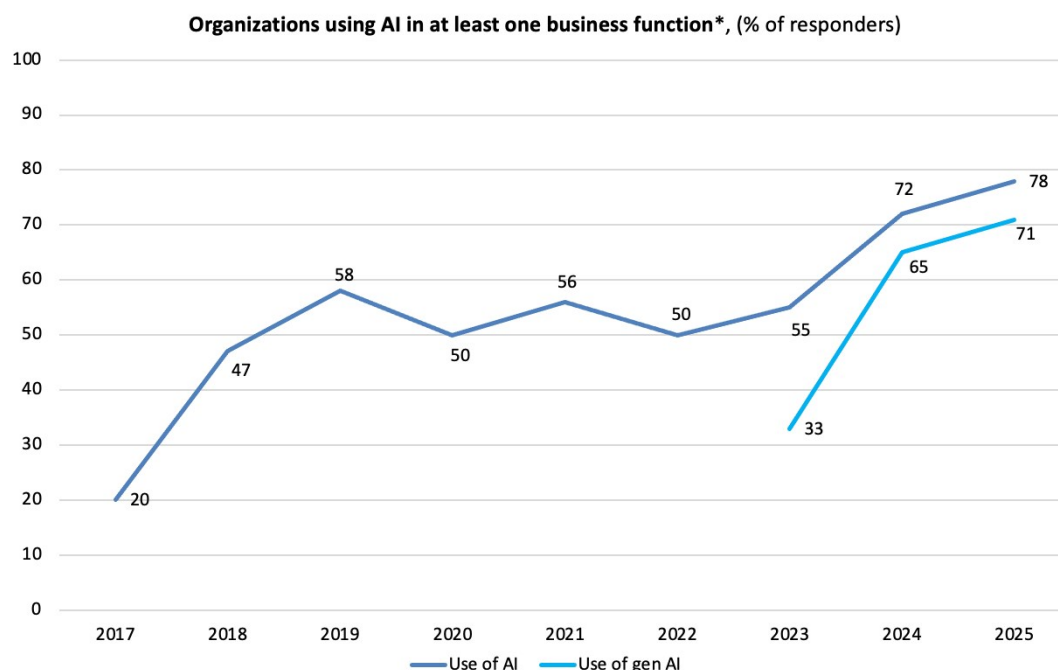
1.1 Background

The accelerating pace of digitalisation is reshaping healthcare and its supporting infrastructures. It creates new opportunities to improve efficiency, safety, and service accessibility. This transformation is also clear in the pharmacy sector, where digital tools and software platforms are central for daily operations. Information technology (IT) supports functions such as dispensing medicines, inventory tracking, and regulatory compliance. A stable digital infrastructure has become critical for modern pharmacy practice (Almeidan, 2024).

Pharmacists play a key role in healthcare and ensure the safe and effective use of medicines. However, their responsibilities are expanding, and administrative tasks are increasing, which created a growing need for efficient technological support (Ogundipe et al., 2025). Help desks play a crucial role in resolving issues related to software usage, onboarding, and system navigation. As pharmacy software grows in scope and functionality, its increasing complexity and the immaturity of supporting IT infrastructure increases the need for timely and reliable technical support (Peltoniemi et al., 2021). Even short disruptions can affect workflows, damage customer trust, and risk patient safety. Help desks play a key role in resolving these disruptions related to software use, onboarding, and system navigation. At the same time, support teams are handling growing volumes of routine and low-complexity queries. Many of these could be managed more efficiently with automation. Reducing this burden is critical for service continuity and for allowing pharmacists to focus on their clinical work (Peltoniemi et al., 2021). Achieving this requires quick access to up-to-date patient information, effective digital tools and consistent digital support solutions (Peltoniemi et al., 2021).

At the same time, artificial intelligence (AI) has emerged as a key driver of digital transformation across sectors (Aldoseri, Al-Khalifa and Hamouda, 2024; Badar et al., 2024; Kraus et al., 2021; Omol, 2024; Varzaru & Bocean, 2024). It is a force that actively drives innovation, creativity, efficiency, and competitiveness across industries (Aldoseri, Al-Khalifa and Hamouda, 2024). A recent survey by McKinsey & Company (2025) found that 78% of organisations already apply AI in some part of their operations. This highlights how AI is increasingly integrated into everyday business activities, as shown in Figure 1. However, given the fast advancement of AI capabilities and the accelerating pace of integration, this figure likely underrepresents the actual scale of adoption. According to Sankar and Sen (2025) conversational AI (CAI) is one of the most promising forms of AI technology. CAI

systems are designed to hold natural interactive dialogues with users. These systems combine multiple AI capabilities and have demonstrated considerable potential in tasks that require contextual understanding and verbal interaction (Feng et al., 2024; Sankar and Sen, 2025). Their applications include customer service automation, knowledge retrieval, and digital assistance (Feng et al., 2024).



The criteria for defining AI use have changed over time. In 2017, an organization was considered to use AI if it applied the technology in a core part of its business or at scale. In 2018 to 2019, the definition shifted to requiring at least one AI capability embedded in business processes or products. Since 2020, AI use has been defined as adopting at least one AI capability in any business function.

Figure 1. Percentage of organisations applying AI in at least one business function (Figure adapted from McKinsey & Company, 2025). Note: This figure shows how the share of organisations using AI in at least one business function has developed from 2017 to 2025. It also reports the adoption of generative AI, which McKinsey began tracking in 2023. The upward trend illustrates the increasing integration of AI technologies across business operations.

1.1.1 Limited CAI adoption in pharmacy IT support

AI adoption in healthcare has grown across both clinical and administrative areas. Examples include risk prediction, diagnostic imaging, and patient triage systems, where AI supports decision-making and improves service efficiency (Zeb et al., 2024). In the pharmacy domain, AI adoption has primarily concentrated on internal processes, with limited application to customer-facing services (Hatzimanolis et al., 2024). Existing applications largely focus on internal process optimisation, such as prescription screening and

inventory control (Hatzimanolis et al., 2024). While these functions are essential, there is a notable research gap concerning the role of AI in enhancing pharmacist-facing or user-oriented technical support functions. The McKinsey & Company (2025) report, shown in Figure 2, points to a clear gap in GenAI adoption in healthcare and pharmacy. Across all industries, 22% of organisations already use GenAI in at least one service operation. In healthcare, pharma, and medical products, however, the figure is only 14%. Meanwhile, sectors such as media and telecom (37%) and technology (30%) are moving ahead much more quickly. This shows that healthcare and pharmacy still have significant untapped potential for GenAI-driven innovation. At the same time, bringing CAI into sensitive healthcare settings raises important questions about reliability, transparency, and ethics. For this reason, it is essential to study how AI tools perform in real pharmacy IT support environments.

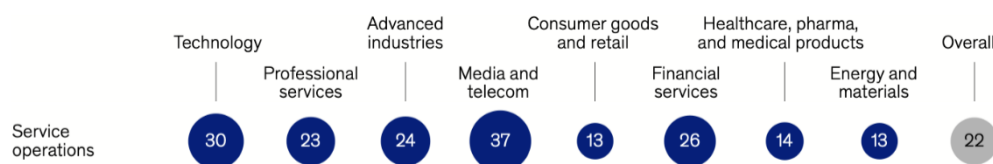


Figure 2. Percentage of GenAI adoption in service operations across different industries (Figure adapted from McKinsey & Company, 2025).

Note: This figure illustrates the share of organisations adopting generative AI in service operations across different industries. The values highlight how adoption levels vary substantially between sectors, with media and telecom showing the highest uptake, while consumer goods, energy, and healthcare including pharmacy IT report comparatively lower adoption rates overall.

Recent advances in CAI, particularly in large language models (LLMs) such as GPT-4 that generate human-like and context-aware responses, have significantly expanded the capabilities of CAI systems (McTear and Ashurkina, 2024). This offers significant potential for improving pharmacy IT support services. AI-powered chatbots have already shown potential to cut repetitive workload and improve response times in pharmacy operations (Adnan et al., 2025). In practice, Finnish pharmacies faced routine technical challenges such as delays caused by searching and signing tasks, immature system integration and occasional technical glitches that disrupted workflows (Peltoniemi et al., 2021). Automating first-line support with CAI could ease staff workload, improve efficiency, and ensure more consistent service. For this reason, pharmacy IT help desks provide a strong and practical context for evaluating the use of LLM-based CAI.

1.2 Topic and scope

This thesis explores how conversational artificial intelligence (CAI) can automate first-line customer support in the pharmacy IT sector. The research is carried out as an empirical case study with Pharmadata Oy, a Finnish company that develops digital solutions for pharmacies. The primary aim is to determine where CAI system powered by LLM reliably replace or complement the workload of human support agents by handling routine and repetitive requests. This includes support topics such as login troubleshooting, onboarding guidance, and navigation within the software platform. These tasks often consume excessive support resources despite being repetitive in nature and relatively low risk.

The motivation for this study stems from the increasing complexity of pharmacy information systems and the rising demand for continuous, high-quality technical support. Pharmacy IT environments must support critical processes such as medication dispensing, inventory tracking, and regulatory compliance (Peltoniemi et al., 2021). All of these require stable and user-friendly digital infrastructure. As these technical systems evolve, support teams receive more and more recurring questions, which limits their ability to address more demanding technical issues.

This study builds on the broader context of digital transformation in enhancing service operations. CAI systems have shown strong potential in many enterprise settings, but their use in pharmacy IT support has been limited (Hatzimanolis et al., 2024). Recent breakthroughs in generative AI have the potential to change this. LLMs are particularly well-suited for tasks that require contextual understanding, quick information retrieval, and natural language interaction (McTear and Ashurkina, 2024). All of these align with the needs of IT support services. This makes pharmacy help desks as a relevant and underutilised domain for the deployment of CAI.

The objective of this thesis is to develop a concept for an AI-based support agent capable of autonomously handling common first-line support requests. A key focus of the research is to assess the CAI's ability to differentiate between queries that are appropriate for automation and those that necessitate human expertise or contextual judgment. The study is based on anonymised historical support data provided by Pharmadata Oy. The data serves as the foundation for analysing real-world queries and evaluating the CAI's decision-making capabilities.

1.2.1 Research questions

To guide the study, the research problem is structured around the following research questions:

1. How accurately can a CAI system generate solutions to selected support requests in real-world pharmacy IT context?

2. How effectively can a CAI system distinguish between cases suitable for automation and cases requiring human escalation?
3. What types of first line in pharmacy IT support requests are suitable for automation using CAI?
4. What system-level structures and requirements are necessary to integrate a CAI-based support agent into existing pharmacy IT workflows?

This thesis focuses specifically on the help desk function, excluding broader clinical applications or full pharmacy management systems. Since the help desk manages the majority of repetitive and predictable requests, it provides a realistic and high-impact entry point for evaluating CAI's potential. Focusing on this area allows a deeper understanding of how AI can improve day-to-day operations in pharmacy software. The study bridges gaps in existing literature by combining insights from enterprise AI deployment, pharmacy IT practice, and applied natural language processing (NLP). The expected outcome is not a deployed AI solution, but a well-grounded concept design. This design will offer a foundation for future development, pilot testing, and decision-making regarding AI adoption in pharmacy IT customer support. Together, these research questions are designed to evaluate CAI performance not only in terms of technical accuracy, but also in terms of operational safety and organisational feasibility.

1.3 Contribution to existing literature

Research on CAI and LLMs has expanded rapidly demonstrating strong capabilities in tasks that require instruction following, summarisation, and procedural guidance (McTear & Ashurkina, 2024; Liu et al., 2023; Wu et al., 2022). In parallel, research on digital transformation highlights that AI systems can enhance operational efficiency and service quality across domains, including healthcare and pharmaceuticals (Almeman, 2024; Badar et al., 2024; Omol, 2024). Despite this progress, the application of CIA in pharmacy IT support remains poorly understood. Existing studies rarely examine how LLM-based CAI performs in real IT support cases that require procedural knowledge, interpretation of technical context, and safe escalation behaviour. Prior research confirms that CAI and LLMs can improve decision support, operational efficiency, and communication accuracy in pharmacy practice (Laymouna et al., 2024; Shin et al., 2024; Wong et al., 2025). However, most of this work focuses on clinical uses (Wong et al., 2025), while the non-clinical domain of pharmacy IT support has received little attention. No prior studies evaluate CAI performance in Finnish pharmacy software ecosystems or analyse which IT support queries can be safely automated without risking misinformation or workflow disruption. This leaves a clear knowledge gap: the suitability, limits and risks of CAI in first-line pharmacy IT support remain largely unknown.

Understanding these gaps is central for advancing both theory and practice. From a scientific perspective, the field still lacks evidence on CAI suitability across task types, the limits of LLM-driven automation, and the mechanisms through which human–AI escalation should operate in regulated environments (Chan et al., 2024; Xiao & Yu, 2025). From a practical standpoint, pharmacy operations rely heavily on stable and accurate IT systems to support medication dispensing, maintain inventory integrity, and meet regulatory obligations (Peltoniemi et al., 2021). Clarifying where CAI can operate reliably directly improves service quality, strengthens workforce efficiency and supports the broader digital transformation in pharmacy IT services.

The contributions of this thesis can be categorised into three primary areas:

i) Real-world insight into CAI performance in pharmacy IT support

The research provides practical evidence of how CAI implemented through Microsoft Copilot environment performs in handling common support queries based on real pharmacy IT service data. The CAI model’s accuracy, task completion, and typical error patterns are evaluated to give a realistic view of its strengths and limitations. This helps extend the academic discussion on CAI performance beyond general use cases into more specialised domains.

ii) A structured framework for identifying tasks suitable for automation

Building on prior research, this thesis offers practical value by presenting a framework for identifying which support queries are suitable for automation. Effective hybrid customer support models rely on clearly defined escalation boundaries that distinguish tasks CAI can safely handle from those requiring human judgment (Li et al., 2024). The approach covers data selection, performance evaluation of the CAI system, and the design of a concept-level support agent structure. This framework provides a systematic method for evaluating CAI in specialised domains such as pharmacy IT, where procedural accuracy, domain expertise, and safe escalation are essential.

iii) Practical relevance for digital development in healthcare

The findings of this study demonstrate how CAI can support pharmacy IT service operations by helping to manage recurring, low-complexity inquiries. Automating such tasks has the potential to ease the burden on support personnel, enhance responsiveness, and improve overall service quality. These benefits align with broader healthcare system goals for digitalisation and more efficient service delivery. The results also highlight the importance of responsible AI use in environments where accuracy, safety, and user trust are critical.

By connecting the potential of CAI with the practical demands of pharmacy IT support, this thesis contributes to a better understanding of how AI technologies can be applied in a responsible way within a complex and regulated environment. This study also encourages continued research on the role of AI in healthcare. It provides a starting point for future work on using AI in technical support systems that must follow strict safety requirements.

1.4 Structure of the thesis

This thesis explores the use of LLM-based CAI agents in automating first-line customer support in pharmacy IT environments. The structure of the thesis has been designed to guide the reader through this investigation in a clear and systematic manner. Each chapter builds upon the previous one, progressing from background and theory to empirical analysis and concept development.

Chapter 2 presents the literature review, which forms the theoretical basis for the study. It begins with digital transformation in the pharmacy sector to situate the research in its broader context. Next, the chapter examines the development of CAI systems and highlights the role of LLMs as the core technology enabling modern AI applications. The chapter also reviews AI agents in service desk environments, human–AI interaction in technical support settings, and the ethical considerations surrounding AI adoption. Together, these sections map existing research, identify knowledge gaps, and clarify why pharmacy IT support offers a relevant yet underexplored setting for CAI deployment.

Chapter 3 introduces the case study of Pharmadata Oy, the partner organisation in this research. It provides an overview of the company’s customer support environment, analyses historical support cases to identify common themes, and maps automatable first-line scenarios. The chapter also explains the setup for testing the LLM-based CAI system, including the criteria used to evaluate model performance in real-world support tasks.

Chapter 4 presents the empirical results of the performance evaluation and analyses how the CAI system behaves across different first-line support tasks in the pharmacy IT context. The chapter examines how the model handles queries and uncertainty, identifies which tasks are suitable for CAI automation and assesses how technical and system-state constraints shape the performance. These findings address the research questions on performance, reliability, task suitability, and escalation. These results form the basis for assessing the system’s strengths and limitations.

Chapter 5 concludes the thesis. It summarises the main findings, discusses their implications for Pharmadata and the pharmacy IT sector, and offers practical recommendations for implementing CAI-based support tools. The chapter closes with a reflection on the study's limitations and suggests directions for future research.

Finally, the thesis includes a complete list of references and appendixes. The appendixes contain supplementary materials such as anonymised support cases, analysis notes, and LLM evaluation records used in the empirical study.

1.5 Research approach and methods

This study uses a qualitative research approach to explore how CAI can help automate first-line IT support in the pharmacy sector. As Dehalwar and Sharma (2023) outline, qualitative research makes it possible to build a detailed understanding of context-specific complex phenomena. It also allows insights to emerge in areas where there is little previous research (Dehalwar & Sharma, 2023). This approach aligns with the exploratory nature of the research problem and the need to investigate CAI behaviour in a regulated, domain-specific environment.

The choice of a qualitative approach is also justified by the complexity of introducing LLM-based CAI agents into a highly regulated, operationally sensitive environment such as pharmacy IT. It requires considering data protection, system integration, support workflows, and user experience (Pasas-Farmer & Jain, 2025). By looking at these aspects together, the study aims to identify both opportunities and limitations for using AI in this domain. The findings seek to support both practical implementation and the broader theoretical understanding of CAI in this field.

The empirical analysis is guided by a working hypothesis that CAI-based automation is most effective in low-risk, single-system, and well-documented support cases, particularly where users seek confirmation or clear procedural guidance. In contrast, multi-system, high-risk, or ambiguous requests are expected to require consistent human oversight. To operationalise this, historical support tickets are classified into task types that differ in complexity and required expertise, and CAI performance is evaluated within each category. Accuracy is used as the primary outcome variable for testing the hypothesis. Additional performance dimensions, such as clarity, helpfulness, structure, terminology and escalation logic are analysed to deepen the interpretation of results. This structure ensures that the hypothesis remains testable while still capturing the multi-dimensional nature of CAI performance in

pharmacy IT support. The detailed implementation of this evaluation framework is described in Chapter 3.

The literature review provides the theoretical foundation of the research. It examines current academic studies on CAI and LLMs in customer service, and AI system integration. Key themes combining these include the capabilities and limitations of CAI, enterprise AI readiness, human-AI collaboration, and the specific constraints of handling sensitive health-related data. The sources include peer-reviewed journals and institutional publications accessed through Aalto University Library and Google Scholar.

The research strategy employed in this thesis is a single-case study. Crowe et al. (2011) define the case study as a research method used to gain a deep and comprehensive understanding of a complex issue within its real-world setting. This is particularly relevant when the boundaries between the phenomenon and context are not clearly evident (Crowe et al., 2011). This approach fits well with this study, as AI integration in the pharmacy sector is closely linked to complex technical systems and strict regulatory requirements. Focusing on a single case makes it possible to study Pharmadata in detail, showing both the wider challenges in the industry and the specific insights from the company's own systems and ways of working.

The main data source is historical customer support cases provided by Pharmadata. These cases document real interactions between pharmacy staff and the service desk. They cover a wide range of support needs, from basic troubleshooting to complex technical issues. This approach integrates the authenticity of real-world data with a systematic analytical framework. Historical support cases offer a comprehensive record of interactions between pharmacy staff and the service desk, capturing both routine and complex support needs. Analysing these cases enables the identification of recurring patterns, the assessment of automation potential, and the evaluation of how effectively an LLM-based CAI system can address different types of support requests. The empirical data is examined using thematic analysis, which helps identify recurring patterns and insights (Clarke & Braun, 2017). Notes from the analysis and the results of the CAI tests are stored systematically, and examples are included in the appendix for transparency. The findings are combined with insights from the literature review to develop a concept model for an AI-assisted support agent that fits the needs of pharmacy IT support. Operational details of the dataset, preprocessing steps and test procedures are presented in Chapter 3.

2 Literature review

2.1 Digital transformation in the pharmacy sector

Omol (2024) describes digital transformation as a profound organisational shift that combines technological innovation with strategic and cultural change. It involves rethinking how organisations function, deliver value to customers, and adapt to a rapidly changing digital environment (Kraus et al., 2021). This transformation extends beyond the adoption of digital tools. It represents a broader process of redefining core business models, operational structures, and internal cultures to align with the realities of a digital-first economy (Omol, 2024).

Digital transformation has emerged as a central driver of innovation across nearly all industries (Egala et al., 2024). Organisations use digital technologies to optimise production processes, automate repetitive tasks, and generate real time insights through data analytics. According to Varzaru and Bocean (2024), these technologies enable companies to achieve greater efficiency, make informed strategic decisions, and create value for customers. However, success depends not only on adopting digital tools, but also on developing an organisational mindset that encourages experimentation and adaptability (Badar et al., 2024).

The pharmacy sector provides a particularly relevant context for examining digital transformation (Hatzimanolis et al., 2024). The shift toward digital technologies in this field can be understood as a gradual process that unfolds in three stages. As Verhoef et al. (2021) explain the first stage is digitisation, which focuses on converting analogue materials into digital formats. In pharmacy practice, this included scanning paper prescriptions, which made storage and retrieval easier but did not fundamentally alter workflows. The second stage, digitalisation involves using digital information to improve efficiency and streamline operations (Verhoef et al., 2021). Software systems for electronic prescriptions, stock control, and labelling improved accuracy and reduced effort but mostly preserved existing practices. The third stage, called digital transformation represents a fundamental shift driven by growing competitive pressures (Verhoef et al., 2021). It not only improves processes but also reshapes how organisations deliver services and create value (Verhoef et al., 2021). At this stage advanced integrated systems are introduced, with AI playing a key role in supporting decision making and data analysis (Hatzimanolis et al., 2024).

Digital transformation is often approached with an excessive focus on technology while the broader business perspective is overlooked. According to

Badar et al. (2025), successful transformation requires rethinking of how organisations operate and strengthening the capabilities needed to support change. It also requires moving beyond existing routines to new, digitally-enabled models (Badar et al. 2025). AI plays a key role in this shift as a strategic driver of innovation and competitiveness (Krakowski, Luger and Raisch, 2023). When combined with automation and process optimisation, AI enables better decision-making and more efficient operations, turning digital transformation into organisational change (Omol, 2024). These developments advance the strategic goals of digital transformation by enabling organisations to adapt quickly and create greater value through technology (Omol, 2024). A major milestone in this development was the release of OpenAI’s ChatGPT in 2022, which accelerated progress in CAI (McTear and Ashurkina, 2024). LLM based CAIs such as ChatGPT are considered disruptive technologies because they challenge established practices and enable new forms of value creation (Coman and Kifor, 2024). By making interactions more natural and intelligent, these systems have reshaped customer communication and knowledge management while also opening new possibilities for innovation in sectors such as healthcare and pharmacy IT support.

Wahl et al. (2024) emphasise the importance of strong governance structures to manage issues such as data privacy, interoperability, and regulatory compliance. In highly regulated sectors such as pharmacy, transformation is not simply a technical process but one that must align with national health policies and legal requirements. Moreover, digital success depends on cultural readiness. As Badar et al. (2025) note, many organisations focus too heavily on tools and neglect the organisational change required to implement them effectively. Leadership commitment, staff engagement, and training are essential to realising the full benefits of digital systems.

2.2 CAI systems

CAI systems are software solutions that allow applications to communicate with users through natural language (Badar et al., 2024). They are often referred to as chatbots or digital assistants and are designed to support tasks such as information retrieval, troubleshooting, and guidance (Laymouna et al., 2024). In this thesis, CAI systems are seen as part of a broader group of agent technologies. AI agents, often described as chatbots or virtual assistants, represent their most common application in practice (Badar et al., 2024).

CAI systems started as simple, rule-based scripts with limited capabilities (Laymouna et al., 2024). With advances in natural language processing (NLP), natural language understanding (NLU) and machine learning, they can now interpret questions, give accurate answers, and link to external

knowledge bases (Alnefaie et al., 2021). This has made them increasingly valuable for customer support where automation helps manage high volumes of routine tasks (Andrade & Tumelero, 2022). In pharmacy IT support, this means they can handle routine queries, guide users through basic processes, and pass more complex cases to human agents when needed (Andrade & Tumelero, 2022). Despite these advancements, traditional CAI models often struggle to interpret context and capture the user's intent or emotional tone accurately (Laymouna et al., 2024). This limitation is particularly evident in situations where empathy and nuanced understanding are essential for effective customer support (Patel & Jain, 2021).

2.2.1 LLMs as the core of modern CAI

LLMs are advanced computational systems designed for NLP (Naveed et al., 2023). They work by predicting the next word in a sequence based on patterns learned from massive amounts of text data (Naveed et al., 2023). This predictive ability allows them to generate coherent and contextually appropriate language, enabling more natural and adaptive CAI systems (Hassija et al., 2023). LLMs represent a major advancement in CAI. Unlike earlier CAI systems based on predefined rules, LLMs use deep learning methods and vast amounts of training data to generate human-like text (Brown et al., 2020).

The performance of LLMs based CAI depends on several factors. Model size and the quality of the training data have a direct impact on accuracy and generalisation (Renze and Guven, 2024). Decoding settings, such as temperature and token limits, also influence the precision and reliability of model outputs (Renze and Guven, 2024). Lower temperature values tend to produce more factual and consistent responses, which is especially important in safety-critical environments like pharmacy IT support. Domain adaptation further improves the relevance and accuracy of LLMs in specialised settings. Fine-tuning on domain-specific data and the use of retrieval-augmented generation (RAG) methods allow models to access approved knowledge sources in real time (Renze and Guven, 2024). This helps reduce common problems such as outdated information or fabricated content, both of which are significant concerns in healthcare contexts (Huang et al., 2025).

The current model landscape includes both proprietary and open-source options, differing in size, data quality, and deployment constraints. Proprietary models such as OpenAI's GPT-4 tend to achieve the best performance across many tasks but usually require enterprise-level infrastructure and strict data governance to meet privacy and regulatory standards (Sterbini and Temperini, 2024). Open-source models, on the other hand, offer more flexibility for customization and domain-specific fine-tuning. However, they often require greater technical expertise for deployment, integration, and ongoing

maintenance (Sterbini and Temperini, 2024). Fine-tuning and retrieval methods make it possible to integrate LLMs into pharmacy IT help desks safely and effectively while reducing risks such as hallucination or the generation of irrelevant information (Huang et al., 2025). When combined with existing IT service management workflows, these models have great potential in automating routine queries, improving consistency, and support escalation processes while adhering to privacy and safety standards.

2.2.2 CAI system architecture and components

The architecture of a CAI system is typically organised into modular layers that enable end-to-end processing of user interactions. These layers include NLU, dialogue management, integration with enterprise systems, and escalation mechanisms. (Renze and Guven, 2024; Singh & Namin, 2025). These elements form a complete pipeline that allows the system to handle user queries while ensuring reliability, safety, and transparency (Renze & Guven, 2024; Singh & Namin, 2025). In pharmacy IT support, the design must also reflect sector-specific needs such as strict workflows and rapid issue resolution (Almeman, 2024).

Figure 3 illustrates the general architecture of a CAI system, showing how user message moves through several modular layers before a final response is generated. The process begins when the user submits a user input, which enters the NLU module. At this stage the system transforms raw text into structured information (Singh & Namin, 2025). This process involves intent detection, entity extraction, and text classification. Intent detection determines what the user aims to accomplish, while entity extraction identifies key details such as user IDs or error codes (Hirschberg & Manning, 2015). Text classification then assigns the request to the appropriate category so it can be routed correctly (Singh & Namin, 2025). Advances in LLMs have made these processes more accurate, allowing the system to deal with complex and sometimes vague inputs more effectively than older rule-based methods (Brown et al., 2020).

The structured representation of the user message is then passed to the Dialogue Management module. This component acts as the decision-making centre of the system, determining the next step in the interaction. Dialogue managers typically combine rule-based policies with adaptive reasoning from LLMs (Singh & Namin, 2025). Rules enforce structure, safety, and compliance, for example, ensuring the CAI does not perform restricted operations or reveal sensitive health information (McTear & Ashurkina, 2024). LM-based reasoning adds flexibility by enabling more context-aware and natural responses (Subramonyam et al., 2024). In regulated environments such as pharmacy IT support, this balance between structure and

adaptability is essential for ensuring reliability, consistency, and trustworthiness. The dialogue manager also connects to the backend component, which interacts with the organisation’s broader IT environment. This may include knowledge bases, IT service management (ITSM) tools, identity and access management systems, or other enterprise applications (Hassija et al., 2023). These backend systems provide verified information such as troubleshooting steps, procedural descriptions, or system status updates, allowing the CAI to give accurate, authorised responses aligned with the organisation’s workflows (Singh & Namin, 2025). In pharmacy IT support, backend integration is especially critical because the CAI must operate within strict workflows and return answers that reflect documented procedures (Almeman, 2024).

After retrieving the necessary information, the system moves to Natural Language Generation (NLG), where a human-readable response is produced. This stage converts structured actions or database outputs into a structured message. NLG ensures that the final response is clear, understandable, and aligned with professional communication norms, especially important in healthcare-related contexts. Finally, the processed output is delivered back to the user through the user interface.

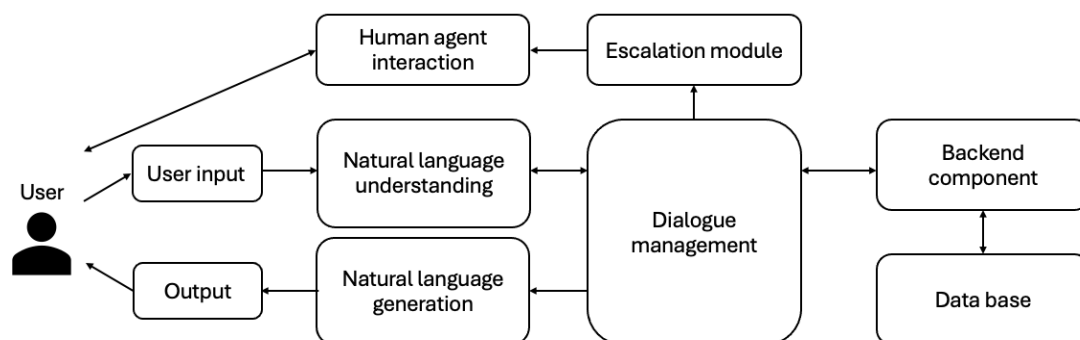


Figure 3. General architecture of a chatbot (Figure adapted from Singh & Namin, 2025). Note: This figure presents the general architecture of a chatbot system. It shows how user input is processed through natural language understanding, dialogue management, and natural language generation. The figure also illustrates how the system connects to backend components and how output is produced. If natural language understanding identifies that escalation boundaries have been exceeded, the query is routed from dialogue management to the escalation module and then to a human agent.

Modern CAI systems also rely on escalation mechanisms to route complex or high-risk requests to human support agents when the system’s confidence is low or the query exceeds its competence (Chan et al., 2024; Xiao & Yu, 2025). In these cases, the conversation is handed over to a human agent. A well-

designed system transfers the full context, including history and extracted details, so the user does not need to repeat information (Patel & Jain, 2021; Singh & Namin, 2025). Xiao and Yu (2024) emphasise that escalation should not be regarded as a failure of automation, but as an essential safeguard within hybrid human–AI systems. Poorly designed escalation undermines user trust, while early and structured handoffs improve both efficiency and customer satisfaction (Xiao & Yu, 2025). In pharmacy IT, this design consideration is particularly important, since unresolved or delayed cases can directly impact the reliability of support and, ultimately, patient care. In CAI architecture, escalation mechanisms are typically implemented as part of the dialogue management layer. Once the dialogue manager interprets the user request and evaluates system confidence, it decides whether to continue the automated flow or hand the interaction to a human agent. For this reason, the escalation component conceptually branches from the dialogue manager, allowing the system to evaluate uncertainty, risk, and task complexity before redirecting the case to human support when necessary (Chan et al., 2024).

2.2.3 Evolving agent interactions

Modern AI systems are designed to evolve continuously rather than operate as static models. They learn from interactions and outcomes by using feedback from both resolved and escalated cases to improve intent detection, expand the knowledge base, and refine dialogue strategies (Singh & Namin, 2025). Over time, this cycle enhances performance and aligns the system more closely with the needs of its users (Badar et al., 2024). Figure 4. illustrates this ongoing cycle of perceiving, deciding, acting, and learning. Interactions form the foundation of this process. As Nguyen (2023) notes, agent systems rely on several interaction types: agent–self, environment–self, agent–agent, environment–environment, and agent–environment to perceive changes, make decisions, and respond effectively.

Agent–self interactions rely on the agent’s internal state, while environment-driven interactions evolve independently of the agent. Agent–agent interactions enable coordination, information sharing, and collective learning, particularly when supported by structured communication mechanisms (Nguyen, 2023). In this context, the environment includes all external elements that interact with the CAI agent. This encompasses human users, support personnel, backend systems, workflows, historical ticket data, and organisational constraints. Human users are therefore a core component of the environment: their queries, corrections, and behavioural patterns influence how the CAI perceives input, how it must act, and how it ultimately learns from each interaction (Nguyen, 2023). For CAI in pharmacy IT support, agent–environment interactions are the most critical. In these interactions, the CAI agent observes user behaviour and adjusts its decisions based on contextual

feedback (Nguyen, 2023). This dynamic loop enables the system to adapt continuously and align more closely with real-world workflows. Figure 4 combines these ideas by linking the linear processing pipeline of CAI architecture with the broader agent learning cycle. When the CAI agent receives a request, it perceives the input, decides on the appropriate action, responds or escalates, and then incorporates the outcome as feedback (Badar et al., 2024). Whether the case was resolved, corrected, or escalated, this feedback strengthens system behaviour over time, supporting safer and more reliable deployment in pharmacy IT environments.

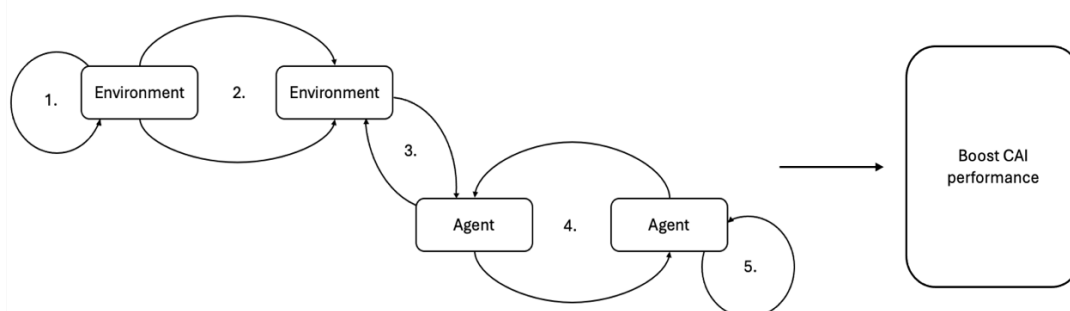


Figure 4. Five interaction types boosting CAI performance. This figure illustrates five types of interactions that contribute to the continuous evolution of conversational AI systems. It shows how performance improves through (1) environment–self interactions, (2) environment–environment interactions, (3) environment–agent interactions, (4) agent–agent interactions, and (5) agent–self interactions. Together, these interaction types support ongoing learning and performance enhancement over time.

2.2.4 Automation and decision-support capabilities

In pharmacy IT support, automation and decision-support capabilities transform CAI systems into intelligent operational assistants that improve efficiency, accuracy, and compliance (Wong et al., 2025). When combined with structured business process management (BPM) principles, these systems form the foundation for AI-enabled service optimisation. Badar et al. (2024) describe BPM as a set of activities that identify, execute, monitor, and continuously improve organisational processes. Automation also evolves into decision support through adaptive learning. Feedback from escalated or resolved cases can retrain the model, improving intent recognition and contextual reasoning over time. This follows Badar et al.’s (2024) adaptive learning and continuous improvement pillar, which promotes iterative enhancement based on AI-derived insights. Over time, the integration of BPM and CAI systems creates a self-optimising, data-driven framework that enhances operational efficiency, supports strategic decision-making, and enables

continuous real-time improvement through AI-driven automation (Badar et al., 2024)

Integrating CAI and BPM allows pharmacy IT environments to shift from reactive, manual service handling to adaptive, process-oriented management. In practical terms CAI systems based on LLMs can classify, prioritise, and route support requests automatically by analysing intent and urgency (Badar et al., 2024). This intelligent triage supports the early BPM stages of process identification and discovery by mapping recurring issues, locating workflow bottlenecks and distributing workload more effectively (Badar et al., 2024). The outcome is reduced manual input, faster case resolution and more reliable workflow systems. These are all critical when managing time-critical incidents such as medication data errors or prescription access problems that can affect patient care.

AI plays a central role in the analytical core of automation. Badar et al. (2024) describe this layer as the extraction of actionable insights from both historical and live data. They frame it as AI-driven process optimisation that refines workflows in real time. In pharmacy IT, this capability supports CAI-driven knowledge retrieval and guided troubleshooting. Using retrieval-augmented generation (RAG), LLMs can access and apply information from standard operating procedures (SOPs), regulatory documents, and verified documentation to produce precise and verifiable instructions to customers (Badar et al., 2024). When a software error pattern occurs, the CAI system can guide the user through structured troubleshooting steps that follow evidence-based instructions. Automation also extends beyond information delivery. Badar et al. (2024) call this seamless automation, where systems complete repetitive tasks such as report drafting, form pre-filling, and checklist creation. In a pharmacy help-desk setting, assisted actions can standardise ticket documentation and generate pre-validated responses. Through this CAI systems become active participants in organisational workflows rather than passive tools, contributing directly to process optimisation and operational reliability.

2.2.5 Technical Challenges and Performance Factors

LLM-based CAI systems offer major benefits but also face significant technical and organisational challenges. Overcoming these is crucial to ensure reliable, safe, and compliance in pharmacy IT, where automation directly supports healthcare operations. Hallucinations and factual accuracy remain a key concern since they are one of the most reported weaknesses (Qu et al., 2025). LLM models can generate content that sounds correct but is inaccurate or unsupported. According to Qu et al. (2025) these hallucinations often stem from misaligned pre-training data, limited fine-tuning, or weak

retrieval grounding. Bias and unfairness are also persistent challenges. Qu et al. (2025) discuss that demographic and contextual biases within the training data can lead to stereotype reinforcement. Instruction non-compliance further threatens reliability. It occurs when the LLM model ignores constraints or responds to adversarial prompts (Qu et al., 2025). Pharmacy systems therefore require strict model alignment and defensive prompting techniques to preserve compliance in regulated environments.

Data quality and poisoning risks present another layer of complexity. Contaminated or manipulated data introduced during training or fine-tuning can create lasting errors that survive through alignment (Qu et al., 2025). Regular dataset audits and secure data-handling protocols are required to prevent contamination. Performance evaluation depends on both computational and human perspectives (Qu et al., 2025). Quantitative measures such as accuracy, relevance and fluency assess technical quality while logical-consistency metrics such as symmetric and negation checks test coherence across similar queries (Qu et al., 2025). Human review remains especially important to confirm alignment in domain-specific use. Model behaviour is shaped by hyperparameter and decoding strategies. Temperature, top-p, and top-k values regulate randomness in output generation (Troshin et al., 2025; Qu et al., 2025). Lower temperatures produce more deterministic results and are generally preferred in factual or regulated domains such as pharmacy IT (Troshin et al., 2025). Maintaining this balance between factual control and adaptability reflects Badar et al.'s (2024) view that automation should remain flexible enough to adapt to new situations while maintaining safety and alignment.

Several mitigation strategies address the most common technical risks. Retrieval-augmented generation grounds responses in verified data sources to reduce hallucination (Qu et al., 2025). Fine-tuning and parameter editing allow domain-specific adaptation without full retraining, while human-in-the-loop validation provides expert oversight (Badar et al., 2024; Qu et al., 2025). Continuous monitoring and adversarial testing help detect privacy leaks or manipulation, while governance frameworks maintain fairness, accountability, and explainability (Qu et al., 2025). From an operational perspective, enterprise deployment, scalability, and compliance introduce further challenges. Pharmacy IT contexts must comply with strict data-privacy regulations and the complexity of integrating AI within existing digital infrastructures (Wong et al., 2025). Persistent issues such as outdated knowledge, bias propagation, and context drift underline the importance of adaptive learning and feedback loops. Following Badar et al. (2024), sustainable AI in technical support must balance automation with resilience. This balance combines algorithmic precision with human oversight, ethical governance, and continuous process optimisation to ensure long-term reliability and trust.

These challenges can be grouped into a set of core performance factors that shape CAI behaviour in regulated environments, summarised in Table 1. Together, they illustrate what performance factors determine the reliability of CAI systems in pharmacy IT support.

Table 1. *Technical and Operational Factors Shaping CAI Performance.* Note: This table summarises key factors that influence the performance and reliability of conversational AI systems. The performance factors include data quality, retrieval grounding, instruction alignment, hyperparameter configurations, human oversight, model drift, and infrastructure integration. The explanations describe how each factor affects accuracy, stability, and safe operation. The references identify the studies that highlight the relevance of each factor for CAI development and deployment.

Performance factor	Explanation	Reference
Data quality, alignment, and domain fit	Poor, biased, or contaminated data weaken accuracy and increase hallucinations.	Badar et al., 2024; Qu et al., 2025
Retrieval grounding and domain adaptation	Verified domain sources and targeted adaptation improve factuality and stability.	Badar et al., 2024; Qu et al., 2025
Instruction following and safety compliance	Strong alignment prevents rule-breaking, unsafe outputs, and adversarial failures.	Qu et al., 2025
Hyperparameter settings and decoding strategy	Generation settings influence determinism, consistency, and factual reliability.	Troshin et al., 2025; Qu et al., 2025
Human oversight and monitoring	Expert review and monitoring detect errors, drift, and manipulation early.	Badar et al., 2024; Qu et al., 2025
Model drift and knowledge freshness	Outdated model knowledge reduces accuracy as systems and workflows evolve.	Badar et al., 2024
Infrastructure integration and scalability	Performance depends on stable integration, low latency, and scalable enterprise deployment.	Wong et al., 2025

2.3 Human-CAI interaction

The adoption of LLMs in technical support demands careful consideration of how humans engage with these systems. Unlike traditional tools that rely on fixed menus or predefined commands, LLM-based CAI systems understand natural language input. This allows users to express their needs in plain text, which lowers the entry barrier and makes the technology feel more intuitive. However, this simplicity also introduces new challenges. Users may not fully understand how the system generates its responses, what its limitations are, or how much they can trust the information it provides (Zafar et al., 2024).

Human- CAI interaction rests on three main principles: transparency, controllability and user agency. Research shows that while CAI makes interaction more flexible and natural, it can also make users feel less in control (Zafar et al., 2024). They often struggle to understand how the system interprets their input, why it responds a certain way, or how to check whether its reasoning is correct. Subramonyam et al. (2024) describe this challenge as part of the “Gulf of Envisioning”, where users find it difficult to translate goals into effective prompts, leading to an uncertainty and trial-and-error behaviour. These problems are magnified in complex tasks, where a single prompt may produce incomplete or low-quality results. Wu et al. (2022) note that while LLMs can flexibly shift roles and act as useful assistants, users struggle to debug unexpected outputs or adapt prompts effectively. In practice, this means LLMs can offload routine tasks and support higher-level decision-making, but only if users trust the system and understand its role. In technical support, this balance is especially critical. Staff rely on timely and accurate answers, yet conversational systems often require trial and error before coming up with the right result. As Wu et al. (2022) and Subramonyam et al. (2024) emphasise, improving adoption depends not only on the technical power of models but also on designing interaction patterns that make them more transparent, controllable, and collaborative.

2.3.1 Prompting and user experience patterns

Prompting is the main interaction mechanism in LLM-based systems. Response quality depends largely on how the request is written (Chang et al., 2024; Wu et al., 2022). Different prompting strategies address this challenge in distinct ways. Table 2 lists some of the most used prompting strategies. They are zero-shot prompting, few-shot prompting, chain-of-thought reasoning and retrieval-augmented generation. Wu et al. (2022) demonstrate that a single prompt often struggles with complex, multi-step tasks. To address this, they propose a chaining approach where each prompt builds on the previous one. This method makes the process clearer and gives users

more control as they can review and adjust each step before moving forward. Prompt chaining is also listed in Table 2. Wu et al. (2022) also introduce prompt templates with clear labels (for example, “Problem:” / “Suggestion:”) and short examples to clarify meaning. In their user study, the chaining method helped participants review and edit results step by step instead of redoing the same prompt, which made collaboration feel smoother and improved the final answers (Wu et al., 2022). Subramonyam et al. (2024) extend this by recommending “intent scaffolding”- templates that guide the user through the task definition process to reduce cognitive load and improve prompt clarity.

Table 2. *Prompting strategies in LLM-based CAI systems. Note: This table presents commonly used prompting techniques for large language models in conversational AI systems. Each strategy outlines how instructions or examples are structured to guide model behaviour, ranging from zero shot and few shot prompting to more advanced methods such as chain of thought, prompt chaining, and retrieval augmented generation. The referenced studies identify where these techniques have been discussed or applied in prior research.*

Technique	Description	Mentioned
Zero-shot prompting	Performing a task without examples and relying only on the instructions provided	Chang et al. (2024); Wu et al. (2022)
Few-shot prompting	The prompt includes a small number of examples to help the model adapt to the expected structure or style	Chang et al. (2024); Wu et al. (2022)
Chain-of-thought	The model is encouraged to produce step-by-step reasoning before giving a final answer, improving accuracy on complex tasks	Patil et al. (2024)
Prompt chaining	A multi-step prompting method where each prompt builds on the previous one	Wu et al. (2022)
Retrieval-augmented generation	Model accesses external knowledge source to improve reliability and context-awareness	Chang et al. (2024); Wu et al. (2022)

While these prompting strategies improve general usability, sector-specific UX patterns are for making CAI tools more usable in pharmacy IT support where precision and accountability are critical (Szymanski et al., 2024). In these environments, staff work within defined standard operating

procedures (SOPs) and structured workflows that benefit from consistent input formats. Prompt-templates can reduce variability by guiding users to frame queries in standardized ways (Li et al., 2024). Interface elements such as “cite SOP” or “show steps” buttons can further strengthen accuracy and accountability by linking outputs to trusted organizational references (Li et al., 2024). To support complex or sensitive cases, built-in handoff functions can route requests from AI to human experts when confidence is low (Li et al., 2024). Manual control is a necessary safeguard that allows users to intervene and correct model outputs mid-process (Subramonyam et al., 2024). This ensures that critical issues receive appropriate oversight. Similarly, confidence badges help users judge when automated results can be trusted, providing an additional layer of transparency (Li et al., 2024).

By combining structured prompting patterns with pharmacy specific UX design, CAI systems can bridge the gap between flexible conversational interaction and the strict procedures required in healthcare. Embedding trust cues and decision-support features directly into the interface builds confidence and reliability, reflecting Norman’s (2013) principles of feedback, visibility, and affordance that define trustworthy human–computer interaction.

2.3.2 Training, onboarding and reducing user burden

Successful adoption of CAI systems in technical support depends not only on system design but also on how users are introduced to the technology. Training and onboarding play a critical role in shaping mental models of how the system works, what it can and cannot do, and how to craft effective prompts (Qu et al., 2025; Wong et al., 2025). Without such guidance, users risk falling into the “capability gap” described by Subramonyam et al. (2024), where they do not know how to express their intentions in ways the model can reliably interpret.

Reducing user burden is particularly important in pharmacy IT support, where staff are under time pressure and cannot afford long trial-and-error interactions (Laymouna et al., 2024). Tools that provide suggested prompts, examples or alternative outputs can help users achieve results more efficiently (Subramonyam et al., 2024). Usability research highlights that system adoption is more likely when the effort required to learn and operate a tool is minimised, and when outputs can be trusted as both accurate and relevant (Laymouna et al., 2024). Transparency features such as step-by-step reasoning explanations or references to official documents increase confidence in system outputs and reduce the risk of over-reliance on AI-generated content (Zafar et al., 2024).

2.4 Challenges and ethical considerations in CAI support

The integration of CAI and LLMs into customer support environments has raised both practical and ethical questions (Zafar et al. 2024). As these systems increasingly handle customer and operational interactions, maintaining trust becomes a central concern. Badar et al. (2024) emphasise that transparency, fairness, accountability, and continuous monitoring are core principles for trustworthy AI. Protecting data privacy, preventing bias, and keeping a human in the loop are essential for trust (Badar et al., 2024). Organisations therefore need clear ethical guidelines and a culture of responsibility so that AI supports both business goals and broader societal benefit.

User trust in CAI systems depends heavily on perceptions of privacy and data security. Leschanowsky et al. (2024) emphasize that concerns about privacy and security remain central to how users perceive CAI and directly impact its trustworthiness. A lack of clarity about data use can undermine confidence and over time, lead to system avoidance or abandonment (Leschanowsky et al., 2024). Similarly, a lack of transparency in how the system generates or explains its outputs can weaken confidence and raise ethical concerns around autonomy and consent (Zafar et al., 2024). Trust can be strengthened through clear disclosure of AI system limitations and verifiable documentation of sources. When users understand how decisions are made and what information supports each output, they are more likely to view the system as reliable (Zafar et al., 2024).

Ethical and security considerations must be built into CAI design from the beginning. Frameworks such as AI Trust, Risk, and Security Management (AI TRiSM) provide structured approaches for addressing data privacy, model governance, and system robustness (Habbal et al., 2024). Core concerns include handling personally identifiable information, preventing malicious prompt injections, and protecting confidentiality (McTear & Ashurkina, 2024). In Europe the regulatory environment for AI in healthcare is becoming increasingly comprehensive (European Commission, 2025). The General Data Protection Regulation (GDPR) ensures the ethical use of sensitive health data while the AI Act gives strict rules for high-risk systems such as medical and pharmaceutical AI applications (European Commission, 2025). These include requirements for transparency, human oversight, and the use of reliable and high-quality data. Together, these measures promote trust, safety, and responsible innovation across the European health ecosystem. For pharmacy IT, these developments underline that CAI must comply with EU principles of transparency, accountability, and data protection to achieve sustainable and trustworthy integration.

The rapid pace of LLM development makes it difficult to keep LLM applications relevant for long. As new versions are released, earlier findings can

quickly become outdated, making it hard to compare performance or reproduce outcomes (Leschanowsky et al., 2024). Chang et al. (2024) also note that fixed evaluation methods can hide weaknesses, since models often learn to memorise benchmarks rather than improve real capability. To address this, more flexible evaluation systems that evolve with each model generation have been proposed as a better way to measure accuracy and safety (Chang et al., 2024). These challenges underline the need for systematic, domain-specific and ethically grounded research. In regulated fields such as pharmacy IT, this means designing evaluation frameworks that consider privacy and compliance from the beginning, ensuring that academic findings translate into the safe and responsible deployment of AI applications.

3 Case study – Pharmadata Oy

Pharmadata Oy is a Finnish health IT company specialising in the development and delivery of pharmacy-specific information systems and digital infrastructure. Founded in 1989 and fully owned by the Finnish Pharmacists' Association (Apteekkariliitto), the company operates as a key enabler of digitalisation in the Finnish pharmacy sector (Pharmadata Oy, 2025). Pharmadata's mission is to serve as a reliable technology partner for community pharmacies by delivering tailored, secure, and user-friendly digital solutions that enhance daily pharmacy operations.

The company's core offerings include pharmacy information systems such as Omapd and pd3, the Apteekkiverkko secure network and a portfolio of additional services including Easymedi, SecureMedi, and Procuero. These systems are complemented by comprehensive Service Desk operations, software training, pharmacy-specific consulting, and automated billing services. Through these services, Pharmadata supports critical pharmacy functions including medication dispensing, inventory management, communications, and compliance with regulatory requirements.

In 2024, Pharmadata reported a revenue of €8.44 million and employed a staff of 44 professionals (Pharmadata Oy, 2025). The company's customer satisfaction rate is high, with 76% of clients reporting they are either very or highly satisfied (Pharmadata Oy, 2025). The company's close connection to the Finnish Pharmacists' Association ensures that its product development remains closely aligned with sector-specific needs and professional standards.

Digitalisation plays a central role in Pharmadata's strategy. As the healthcare and pharmacy sectors undergo increasing digital transformation (Kraus et al., 2021), Pharmadata aims to lead the field by developing innovative IT solutions that improve efficiency, data security, and ultimately enhance the quality of pharmaceutical services across Finland. Alongside product development, Pharmadata is working to improve the efficiency and availability of its customer support as digital service demands increase. The company's long-standing expertise and domain knowledge provide a strong foundation for this development. In this context, the evaluation of an LLM-based CAI system in first-line customer support forms part of Pharmadata's ongoing, proactive development efforts. The aim is to streamline support processes and improve the efficiency, availability, and accessibility of customer support services, making it easier for customers to obtain even faster assistance.

3.1 Introduction to case study

This case study looks at how LLM-based CAI systems could help automate first-line support tasks in Pharmadata's customer service environment. Pharmadata was selected as the case study company because it plays a central role in the digital transformation of pharmacy IT. Reliable technical support is essential to ensure uninterrupted pharmacy operations. At the same time, digitalization is driving the shift from traditional customer service practices toward AI-assisted solutions, helping Pharmadata meet the demands of the digital era and stay competitive.

This research uses historical customer support cases as its primary data source rather than interviews. The dataset contains 5,855 unique support tickets recorded between 1st of August and 15th of September 2025. It is important to note that this ticket volume reflects Pharmadata's support process rather than the number of distinct problem situations. Support tickets are closed after each response, and if further actions or clarification are required, a new ticket is opened for the same underlying issue. As a result, a single customer case may be represented by multiple consecutive tickets. This practice supports clear case tracking and process transparency, but it also increases the total number of recorded tickets. The selected timeframe of 1st of August and 15th of September 2025 represents a typical operational period for Pharmadata's Service Desk, with stable staffing levels and regular system usage following summer holidays. These customer support cases document real interactions between pharmacy staff and Pharmadata's Service Desk, ranging from simple questions about system features to complex troubleshooting requests. Analysing actual support cases gives a realistic picture of the challenges faced in daily operations and shows where AI might offer the most value. It also helps identify patterns in support requests, such as recurring technical issues or frequently asked questions, which can inform the design of AI-driven solutions. By grounding the analysis in real-world data, the study ensures that proposed AI applications address practical needs rather than hypothetical scenarios.

3.1.1 Analysis design

The analysis is carried out in two main stages, with all steps carefully documented to ensure transparency and reproducibility. In the first stage, anonymized support tickets are systematically sorted by topic, complexity, and resolution method. Detailed notes are taken at each stage to capture patterns, exceptions, and decisions made during the sorting process. This structured approach makes it possible to identify repetitive first-line requests that are strong candidates for automation. The categorised inquiries serve as the foundation for training the Copilot CAI-model within Pharmadata's secure Microsoft 365 environment. Based on this structured data, Copilot learns to

identify recurring question patterns and to recognise situations that require escalation to a human support agent. During this phase, clear escalation prompts, and decision rules are also determined. These provide a framework for later evaluating the model's judgment, accuracy, and reliability.

In the second stage, the trained Copilot model is tested and analysed using Pharmadata's secure Microsoft Copilot environment, which provides access to LLM-model capabilities within a protected and compliant setting. The model is evaluated using a set of sample cases and support questions that simulate realistic first-line scenarios. The goal is to assess how effectively Copilot interprets requests, generates accurate and understandable responses, and applies the defined escalation criteria. Each interaction between the model and the anonymized support cases is logged, including the model's responses, decision-making steps, and any points where the system chooses to escalate the case to a human. These notes form a comprehensive record of the model's behaviour, enabling both qualitative and quantitative assessment of its performance. After testing, results of each test conversation are assessed using a structured questionnaire (Appendix 1, Testing questionnaire). The form gathers user feedback on clarity, accuracy, and terminology use. It also measures whether AI responses save time, require human corrections, or correctly escalate cases when necessary. Responses are rated on a five-point scale ranging from "completely disagree to completely agree" providing a systematic framework for analysing user perceptions and model performance.

Insights from the literature review shape the analysis design. Research on CAI architectures and LLM capabilities informs how the model's accuracy, relevance, and context awareness are evaluated. At the same time, studies on human-AI interaction and AI ethics guide the handling of sensitive information. The analysis prioritises usability, reliability, and data privacy throughout the process.

The case study focuses on two main objectives:

1. Identify which support request themes that can be reliably automated using LLM-based CAI.
2. Assess the performance and professional quality of AI-generated responses.

Ultimately, the results aim to show both the potential and the limitations of using LLM based CAI in this setting. This research also provides practical guidance for designing an AI support agent that meets Pharmadata's operational needs while aligning with the ethical and technical requirements of the pharmacy sector.

3.2 Overview of Pharmadata’s customer support environment

Pharmadata’s customer support plays a key role in keeping Pharmadata’s IT systems running across Finland. The team assists pharmacies with technical incidents, software updates, user guidance, and configuration issues related to the company’s main platforms: pd3, Omapd, and Easymedi. In addition to its core platforms, the support team manages various supplementary systems that play an essential role in pharmacy workflows. Support requests arrive through telephone and email, which are the primary channels for customer communication. Customer support operates only during office hours, and no 24/7 service is currently available. Both channels feed into the internal ticketing system once an agent logs a case. The process is manual, and the support personnel write each ticket and customer communication. Manual entry allows individual review, but it adds administrative work and increases variation in how tickets are documented. After a ticket is created, it is assigned a priority level from low, medium, high, or critical based on urgency and impact of the issue. Most tickets (96.7%) fall into the medium category, representing standard user inquiries and non-urgent technical issues. High and critical cases together account for only about 2.9% of all tickets, yet they consume a big share of resources due to their impact on critical pharmacy operations. Major priority tickets (0.4%) represent complex system-level issues usually requiring cross-team collaboration. The distribution of tickets and their corresponding resolution times are summarised in Table 3.

Table 3. *Distribution of historical support tickets by priority level. Note: Summarization of the distribution of Pharmadata’s historical support tickets across four priority categories between 1 August and 15 September 2025. The results show that most tickets were classified as medium priority, while critical, high, and major tickets accounted for only a small share of total requests.*

Priority level	Share of tickets (%)
Critical	1.0
High	1.9
Major	0.4
Medium	96.7

Pharmadata also maintains two digital self-service portals, Neuvo and Oma-Neuvo, which serve as the company’s main knowledge base. Both are

accessible through the secure Apteekkiverkko network and provide step-by-step guides, troubleshooting instructions, and answers to frequently asked questions. The portals are designed to help pharmacy staff find solutions independently and reduce routine inquiries to the Service Desk. In practice, Neuvo and OmaNeuvo are designed to support the customer support team's work by freeing time for more technical and high-priority cases. In addition to the online portals, Pharmadata regularly organises webinars and in-person training sessions for pharmacies using its systems. These sessions cover system updates, new functionalities, and best practices to strengthen users' digital competence.

Overall Pharmadata's customer support is structured and reliable but labour-intensive. It depends heavily on manual documentation and individual expertise. The lack of automation in ticket creation and classification points to a clear area for improvement. This environment provides the basis for exploring how CAI and LLMs could make first-line support faster, more consistent, and easier to manage.

3.2.1 Support demand by system platform

The analysis of Pharmadata's customer support tickets provides a detailed view of where Pharmadata's Service Desk resources are most heavily utilised. Categorizing the tickets by system platform makes it possible to identify where technical incidents, usability issues, and configuration problems most frequently originate. Understanding this distribution helps reveal how digital tools, workflows, and support structures interact in pharmacy environments (Wong et al., 2025).

Figure 5 shows how the tickets were distributed across Pharmadata's main service platforms. Almost half of all cases (45%) were linked to the pd3 system, followed by the Omapd portal (22%) and devices (10%). Smaller shares concerned Microsoft 365 services (4%), member services (3%), and the Pharmacy Network (Apteekkiverkko) (1%). The "Others" category (15%) included smaller platforms such as Holvi, Procuero, SecureMedi, handheld devices, and Proselecta. The dominance of pd3 and Omapd-related tickets highlights their importance in daily pharmacy work. These systems handle key functions such as dispensing, billing, inventory management, and prescription processing. Their high-ticket volumes indicate that even small technical or procedural issues can have wide-ranging operational consequences (Wong et al., 2025). This reflects earlier findings that pharmacists spend around 15% of their time navigating technical systems (Wong et al., 2025). AI tools could help shift this effort towards more meaningful clinical work.

Device-related requests also demonstrate how much hardware reliability directly affects workflow efficiency. Printers, barcode scanners, and workstations are critical components of pharmacy operations, and their failures can interrupt dispensing or stock control. This has also been noted in automation-efficiency research (Badar et al., 2024). The smaller yet persistent share of tickets related to Microsoft 365 and member services shows the importance of background systems that enable secure access, authentication, and communication across the pharmacy network. Similar patterns have been observed in other healthcare IT contexts, where user support needs often stem from a combination of system complexity, process integration, and human factors (Laymouna et al., 2024; Wong et al., 2025).

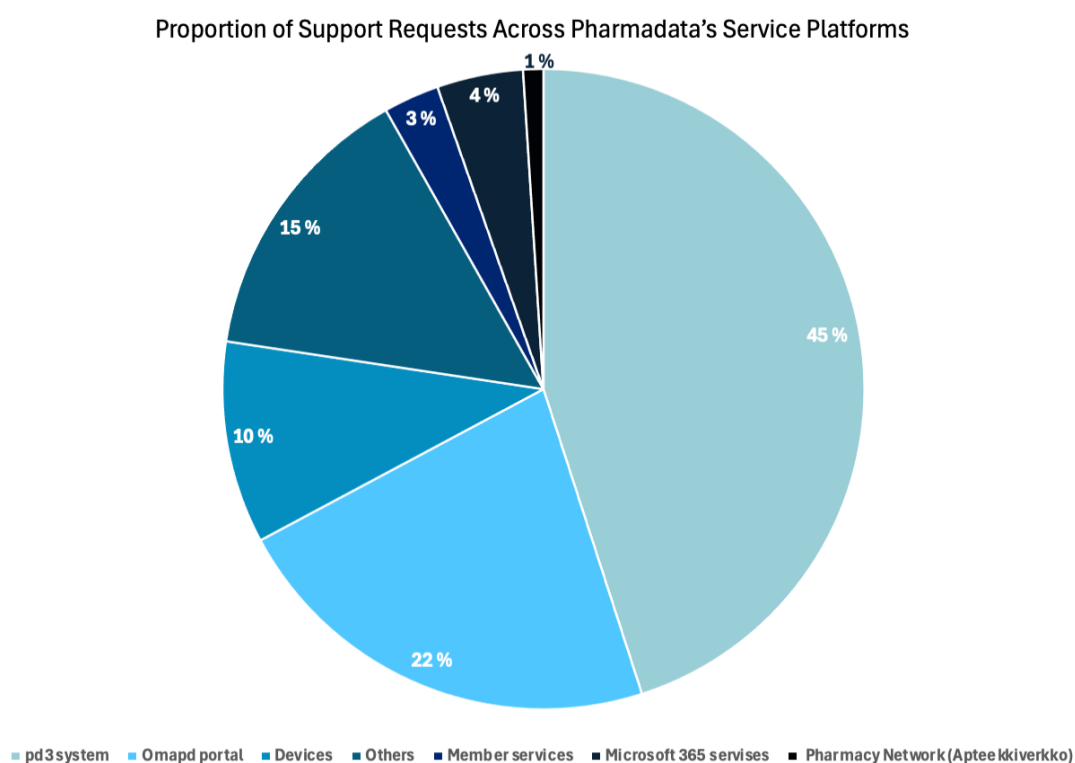


Figure 5. Proportion of support requests across Pharmadata's service platforms. Note: This figure summarises the distribution of support requests across Pharmadata's service platforms. The results show that the pd3 system and Omapd portal account for most requests, while devices, member services, Microsoft 365 services, the pharmacy network, and other categories represent smaller shares of overall support demand.

3.2.2 Support demand by functional area

Analysing the same dataset by functional area reveals which types of tasks generate the greatest need for support. Figure 6 presents the breakdown of tickets across major service categories. The largest shares are miscellaneous

and other (26%), billing (19%), prescription handling (11%) and device issues (11%). Together, these categories represent the areas where pharmacy staff most frequently seek guidance or technical support (Laymouna et al., 2024).

A particularly notable finding in Figure 6 is the large share of tickets in the category miscellaneous. This category includes cases that do not fit into a specific other category, involve several systems at once or come from a smaller service area that are not large enough to form their own category. The high percentage suggests that many of these problems are cross-system issues, where users experience errors at the boundaries between different applications. For example, a failed data transfer, delayed update, or unclear error message can make it difficult for the user to know where the issue started. Another reason for the large share may be the manual nature of ticket logging, where each case is described and categorised by hand. When agents interpret or label tickets differently, some cases end up grouped as “other.” This type of categorisation challenge is common in complex digital environments, especially in healthcare, where overlapping systems and inconsistent documentation make classification more difficult (Laymouna et al., 2024). The size of the miscellaneous category highlights the need for clearer ticket categories and consistent logging practices all of which could make support more efficient.

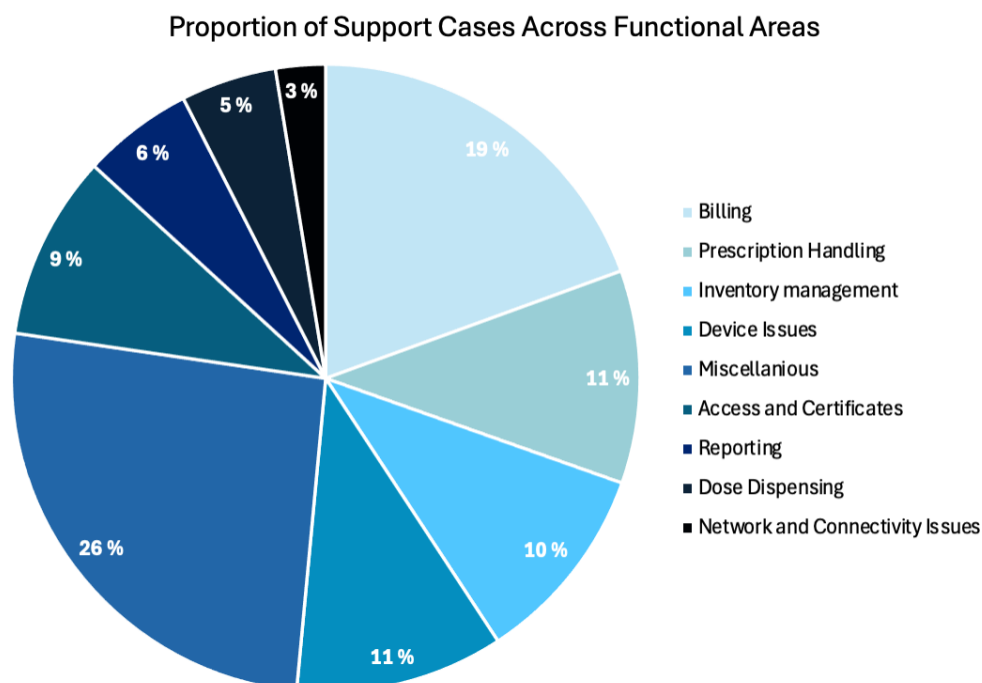


Figure 6. Proportion of support request cases across Pharmadata’s functional service areas. Note: Illustration of how support request cases are distributed across Pharmadata’s functional service areas. The results

indicate that billing and prescription handling account for the largest share of cases, while device issues, reporting, dose dispensing, and network or connectivity issues represent smaller portions of the overall support volume.

Other big groups of support tickets such as concentration in billing and prescription handling reflect the complexity and regulatory precision required in pharmacy IT. These areas cover a large part of daily pharmacy work and depend strongly on accurate data and correct system rules. Billing cases often involve invoice corrections, missing transactions, or validation errors. All of which can occur when data does not align with system or pricing logic. Prescription-related tickets frequently include rejected electronic prescriptions, incomplete renewals, or medication code errors. These demonstrate how strict validation and compliance rules can lead to repetitive, rule-based error types (Laymouna et al., 2024). Alongside these software-related issues, a large share of tickets concern devices such as barcode scanners and workstations. These hardware components are deeply integrated into daily workflows for dispensing and stock control, and any malfunctioning can disrupt operations or delay service. The share of device-related cases underscores how closely software reliability is linked to hardware performance in pharmacy environments (Badar et al., 2024; Omol, 2024).

Smaller but important categories include inventory management (10%), access and certificates (9%), reporting (6%), dose dispensing (5%) and network or connectivity problems (3%). Inventory management tickets point to ongoing challenges in synchronising product data and stock levels between systems. Discrepancies in counts or delays in updates can result in confusion during ordering and restocking. Tickets related to access and certificates illustrate recurring challenges in authentication and security management, particularly during password resets, certificate renewals, or after system updates. Reporting and dose dispensing are smaller in ticket amounts but operationally significant. Reporting issues often concern the generation or interpretation of compliance and management data, while dose dispensing problems directly affect medication safety (Omol, 2024). Network and connectivity problems though few, can have disproportionate effects by temporarily disabling access to all other systems (Omol, 2024).

3.2.3 Overview of structural patterns in support demand

The two platform-level and the functional area analysis reveal that most of Pharmadata's customer support needs come from a limited set of systems and functions that are fundamental to pharmacy operations. The data reveals that technical and procedural questions are a regular part of daily work. Laymouna et al. (2024) found that unclear documentation and routine usability

issues account for most support contacts, especially when systems are complex and tightly integrated. This suggests that many requests arise from everyday tasks rather than major failures, highlighting the importance of user experience, documentation quality, and workflow clarity in shaping the Service Desk workload. Dzindolet et al. (2003) also suggest that pharmacies rely heavily on support not only for resolving faults but also for confirming correct procedures, particularly in billing and prescription processing. The dominance of pd3 and Omapd tickets underscores their importance as critical infrastructure, while device and miscellaneous cases highlight the interdependence of hardware, software, and user knowledge. Understanding these recurring themes provides a basis for identifying repetitive first-line scenarios that could be automated.

3.3 Determining escalation boundaries between CAI and human support agents

In a hybrid customer support model, it is essential to define clear boundaries between tasks that CAI can manage and those that require human expertise (Li et al., 2024). Escalation boundaries describe this division of responsibility. They help ensure that automation improves efficiency while maintaining accuracy, safety, and user trust (Li et al., 2024). Hybrid support systems that combine AI automation with human oversight have been shown to achieve the best balance between reliability and responsiveness, provided that the boundaries of human intervention are well defined (Li et al., 2024). In Pharamadata's Service Desk environment, the ticket analysis presented in Chapter 3.3 revealed that a large proportion of customer requests are predictable and repetitive, which makes them promising candidates for CAI automation. However, some cases involve complex multi-system dependencies that require professional human judgment.

Tasks suitable for CAI automation share three key characteristics: they are low in complexity, minimal in risk, and high in contextual clarity (Li et al., 2024; Wu et al., 2022). The detailed escalation criteria are presented in Table 4, which outlines the decision logic applied to this study. In practice if a request meets even one of the escalation conditions described in Table 4, it should be directed to a human support agent. This strict and conservative approach ensures that automation is applied only when all key factors fall within the safe operational range. This policy prioritises reliability and compliance over automation coverage aligning best with practices in healthcare and enterprise AI deployment.

Low-complexity cases are typically confined to a single system and follow routine workflows with predictable outcomes. Examples include password resets through verified self-service portals, assistance with email

configuration, or guidance through standard troubleshooting procedures. In these cases, the CAI can provide accurate and consistent support without the need for contextual reasoning. Risk level is another critical factor. Tasks involving little or no potential harm to compliance, data integrity, or patient safety can be automated with limited supervision. Typical examples include updating user contact details, retrieving non-sensitive information reports, or confirming standard process steps such as order status. Focusing automation efforts on these low-risk interactions allows the system to improve operational efficiency while avoiding scenarios where incorrect responses could have broader implications (Patil et al., 2024). Equally important is the clarity of context. When the user's intent and the required actions are clearly defined, the CAI can operate with a high degree of confidence. These cases often involve situations where users already have the capability to complete the task but lack awareness of the correct procedure or cannot easily locate the relevant instructions. These "information-seeking" interactions represent the most promising category for CAI-driven automation. They rely on procedural knowledge rather than decision-making which in a highly regulated environment should be left to humans. The success of automating these types of cases will be tested in this study.

Escalation is needed when a request is too complex for the CAI to handle or when making a mistake could have serious consequences. Research on healthcare and CAI systems emphasises that well-defined escalation triggers protect both efficiency and user trust by addressing complexity, risk, and clarity (Wu et al., 2022; Dzindolet et al., 2003). Highly complex cases often span multiple systems or require reasoning across organisational boundaries, making automated decision-making unreliable. Similarly high-risk cases and those involving sensitive data or patient safety demand human intervention to ensure accountability and compliance (Laymouna et al. 2024). Complex clarity is the third reason for escalation. When a user's request is unspecified or the CAI's confidence in its interpretation is low, escalation prevents potential miscommunication and service errors. Typical examples include requests to remove email accounts tied to regulated workflows, the creation of custom user roles with non-standard permissions, or any case where the CAI detects conflicting inputs or potential safety implications.

Table 4. *Escalation Criteria for CAI to Human support in Pharmadata's Service desk. Note: The table summarises the criteria used to determine whether a CAI system can resolve a support request independently or whether escalation to human support is required. The criteria address task complexity, risk level, contextual clarity, data source availability, and system permissions, reflecting operational safety, regulatory constraints, and the need for reliable context understanding in pharmacy IT support.*

Criteria	CAI can handle independently	Escalate to human support	Rationale
Complexity	Single-system requests or multi-step tasks with documented instructions (e.g., log out/relogin, restart device)	Multi-system cases without clear procedural guidance or requiring troubleshooting beyond documented steps	Multi-system dependencies require contextual reasoning beyond the CAI's capabilities. However, with well-documented instructions such cases can be automated safely (Wu et al., 2022; Li et al., 2024)
Risk level	Tasks with no impact on compliance, privacy, or safety (e.g., updating contact info, retrieving non-sensitive data)	Any request involving regulated data, patient safety, or irreversible changes (e.g., deleting prescription data)	High-risk operations require human accountability and auditability (Patil et al., 2024; Laymouna et al., 2024; Dzindolet et al., 2003). Added condition: CAI only acts where user can safely perform the action themselves
Clarity of context	Clearly defined, information-seeking tasks where user intent and procedure are explicit (e.g., "How do I reset my password?")	Clearly defined, information-seeking tasks where user intent and procedure are explicit (e.g., "How do I reset my password?")	Expanded CAI scope to include instruction-based resolutions (Wu et al., 2022; Li et al., 2024).
Data source availability	Cases where verified instructions exist in Neuvo or ticket history	Requests with no documented solution or requiring creative problem-solving	CAI must not generate new content. Answers must come from verified internal data (Li et al., 2024; Leschanowsky et al., 2024; Wu et al., 2022).
System Permission	Read-only or advisory actions and user has required permissions	Write-access or configuration changes beyond user capability	Prevents unauthorised or irreversible actions by AI (Patil et al., 2024; Li et al., 2024).

Establishing these boundaries is essential to maintaining credibility of hybrid support systems. Research on AI transparency and control shows that adding guardrails and checkpoints lets humans' step in when the AI faces uncertain or critical situations (Wu et al., 2022; Leschanowsky et al., 2024). In Pharamadata's operational context, this approach provides a structured method for defining safe automation limits.

3.4 Test implementation: LLM response evaluation

The purpose of the evaluation study is to determine how effectively Microsoft Copilot can support first-line customer service tasks in the pharmacy IT environment. The goal is to determine whether Copilot can generate clear, accurate, and helpful responses to real support queries derived from Pharmadata's service desk operations. By analysing the model's behaviour in realistic scenarios, the study assesses whether Copilot can function as a reliable foundation for an intelligent, LLM-based assistant in a regulated healthcare technology setting.

The evaluation is conducted inside Pharmadata's secure Microsoft 365 environment to ensure full compliance with confidentiality and data protection policies. This controlled setup allows the testing process to reflect authentic service desk workflows while maintaining controlled and safe operating conditions.

3.4.1 Evaluation framework

The evaluation follows a structured framework that collects data on six criteria scoring system that represents key qualities expected from a first-line support agent: clarity, accuracy, helpfulness, structural coherence, terminology use, and escalation logic. These criteria form the basis of the assessment form used throughout the study (Appendix 1). The framework ensures that Copilot's answers are evaluated consistently and provides a systematic way to compare performance across different query types. Each test case is first examined qualitatively. Appendix 1 includes qualitative observations of strengths and weaknesses, and these notes serve as the basis for the numerical scoring. They help determine how well the CAI agent performed in relation to each evaluation criterion. By completing the qualitative assessment before scoring, the evaluation ensures that numerical ratings reflect a consistent interpretation of the model's behaviour and remain grounded in the detailed performance characteristics identified during testing.

Each response is rated using a five-point Likert scale (1 = strongly disagree, 5 = strongly agree), allowing for systematic comparison between different responses (Shirahama et al., 2024; Huang et al., 2019). The evaluation framework was inspired by usability principles and exploratory software testing methods (Shirahama et al., 2024). The evaluation framework builds on recent research showing that the Likert scale is a reliable and efficient way to measure user perceptions, though it can be less sensitive to subtle differences in AI-generated responses (Shirahama et al., 2024). Although no external test users were involved, the structured self-evaluation process provided an

objective way to assess Copilot’s linguistic and functional performance while ensuring reproducibility.

The scale is interpreted as follows:

- 1 – Strongly disagree: Response is incorrect, unclear, or unusable.
- 2 – Disagree: Response is partially correct but needs significant revision.
- 3 – Neutral: Response is usable but incomplete or inconsistent.
- 4 – Agree: Response is mostly correct, clear, and functional.
- 5 – Strongly agree: Response fully meets the criterion without need for correction.

3.4.2 Methods

The evaluation sessions are completed individually using a systematic step-by-step procedure designed to replicate real first-line support interactions. All test sessions are performed individually using Pharmadata’s secure Copilot Studio environment. Before the testing sessions begin, the CAI agent is configured for the evaluation task. The preparation process includes three key steps. First, the agent is instructed on its role and expected behaviour, including acting as a first-line support assistant and providing responses grounded in Pharmadata’s established support practices. Second, to ensure safe and accurate behaviour, the agent is restricted to use only Pharmadata’s anonymised historical service ticket data as its knowledge source. It may draw solely on previous customer queries and their resolved answers and must not generate new information. This constraint ensures that all responses remain grounded in validated internal expertise. Third, the agent is instructed to always respond in the same language as the user’s input prompt, enforcing linguistic consistency and enabling realistic evaluation of Finnish-language support interactions. In addition, the agent is given explicit behavioural constraints based on the escalation boundaries defined in Table 4, instructing it to escalate any query that exceeds these criteria.

Once the agent is configured, the evaluation proceeds with individual testing for each query. Each test query is entered manually into the *Test your agent*-interface, allowing interaction with the model in conditions that mirror real end-user experiences. All prompts were written in Finnish, reflecting the primary working language of pharmacy staff. Each interaction is documented including a short summary the input prompt, Copilot’s output, qualitative notes describing strengths, number of prompts and weaknesses of the answers and numerical scores for each criteria evaluated in Appendix 1.

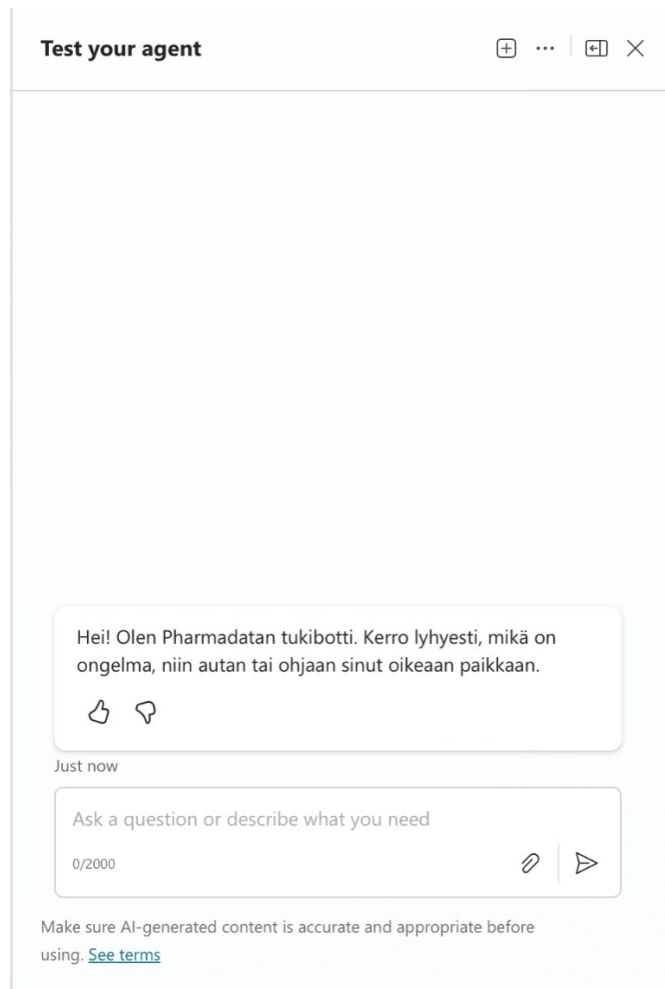


Figure 7. Copilot Studio CAI agent test box example. Note: This figure shows the test interface for the CAI agent in Microsoft Copilot Studio. The example includes a Finnish system message that introduces the support bot to the user. In English, this means: “Hi! I am Pharmadata’s support bot. Briefly describe your issue, and I will help you or direct you to the right place.” The test interface allows for testing of the agent configured, and review of how the agent interprets user input and generates responses during testing.

The Copilot Studio test agent used in the evaluation relies on ChatGPT-5–based LLM capabilities, which enable advanced reasoning and contextual interpretation. Accordingly, the evaluation examines not only the clarity and coherence of Copilot’s responses but also how effectively the model retrieves information and aligns its answers with Pharmadata’s historical support solutions. This includes assessing whether the model can correctly identify the relevant historical cases, reproduce the essential troubleshooting steps, and apply appropriate domain reasoning. Equally important is Copilot’s ability to recognise when a query exceeds its operational scope. Using the escalation logic defined in Table 4, the evaluation determines whether the model

escalates such cases appropriately and in a manner consistent with historical agent behaviour.

3.4.3 Test data and case selection

A total of 100 test queries is created based on actual data from Pharmadata's customer support tickets. The queries are designed to mirror the structure, language, and content of real user requests while ensuring full anonymisation. Each test case is conducted from authentic support interactions but reworded to avoid identical phrasing with training data. This approach maintains realism while protecting customer confidentiality.

The selected queries cover all main categories identified in Chapter 3.2.1, ensuring that the evaluation reflects the full scope of first-line support tasks. These categories include typical issues such as login failures, device errors, medication record inconsistencies, and system access requests. Each category is represented by several examples to capture both straightforward and complex situations. The test set is designed so that queries remain highly similar in tone and intent to the original customer messages. This helps replicate the language and problem-solving context faced by Pharmadata's service desk, allowing the evaluation to measure Copilot's ability to interpret real-world user input accurately.

To ensure comprehensive coverage the 100 test queries are divided into functional areas that reflect Pharmadata's operational priorities. The categorisation aligns with historical ticket data and is visualised in Figure 8. This distribution ensures that the evaluation covers both routine and complex cases across all major service desk functions. Most of the functional category includes tasks that are suitable for automation as well as tasks that, according to the escalation criteria in Chapter 3.3, require transfer to a human agent. The only one not including a test query requiring escalation is Device issues. The distribution between automatable queries and queries requiring escalation is shown in Figure 9. By using realistic yet anonymised queries across all support categories, the evaluation provides a comprehensive and representative foundation for testing. It allows analysis of how well Copilot recognises intent, applies appropriate terminology, and adapts to varying query complexity, while also identifying where escalation to a human agent is still required.

Distribution of test queries across functional areas

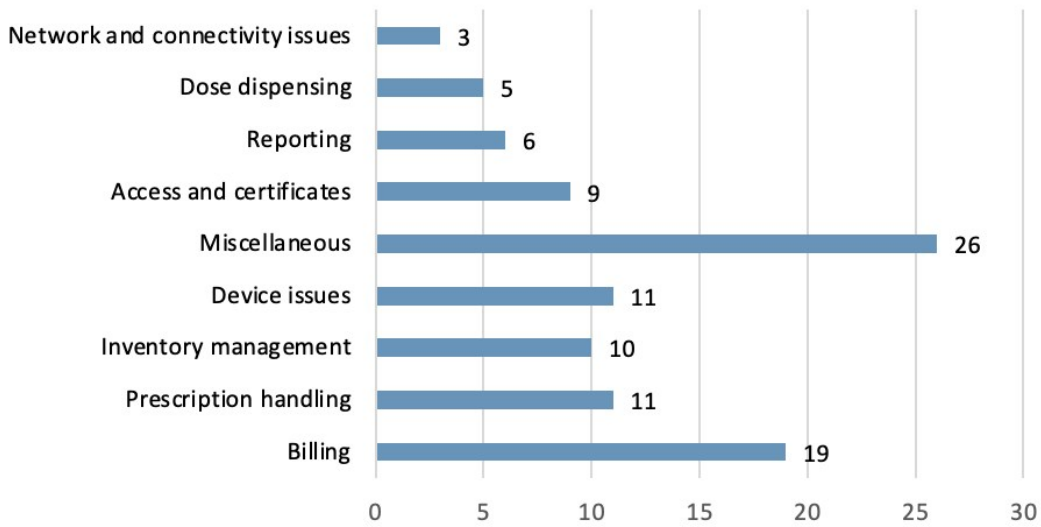


Figure 8. Distribution of test queries across Pharmadata’s functional areas. Note: Illustration of how the test queries used in the evaluation were distributed across Pharmadata’s functional service areas. The largest number of queries concerned miscellaneous issues and billing, while areas such as reporting, dose dispensing, and network or connectivity issues were represented by fewer test cases. The distribution reflects the percentage share of historical queries each category contributed to the overall training data.

Distribution of test queries requiring escalation vs automation

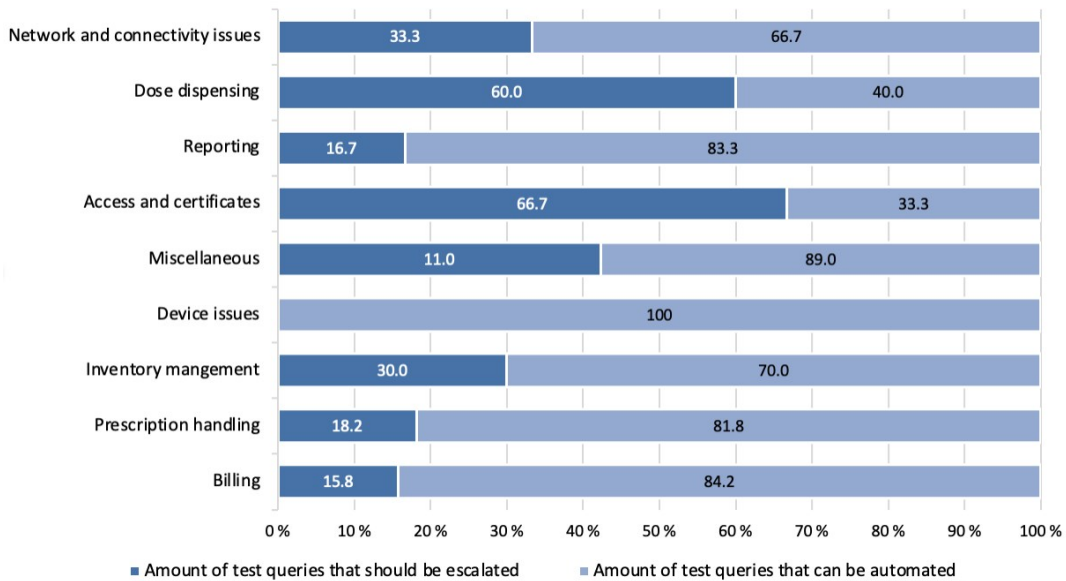


Figure 9. Distribution of test queries requiring escalation versus automation. Note: This figure shows how the test queries were divided between cases that could be automated by the CAI agent and cases that required

escalation to human support. The proportions vary across functional service areas. Tasks involving device issues, dose dispensing, and network or connectivity issues had higher escalation rates, whereas categories such as miscellaneous queries and billing showed greater potential for automation.

3.4.4 Data recording and analysis

The data recording process begins immediately after each test interaction. Collected evaluation scores are transferred to a central spreadsheet that organises the dataset in a consistent and analysable format. Each row represents one test case and includes the anonymised query identifier, the model's response, qualitative notes describing strengths, errors, and reasoning patterns, and the numerical scores assigned to each evaluation criterion. This structure ensures full traceability and makes it possible to link the model's observable behaviour with its quantitative outcomes.

Once all cases are documented, the numerical scores are summarized to generate descriptive statistics for each evaluation criterion and functional area. These summaries allow comparison across task types, highlight performance patterns, and show where the model performs consistently well or poorly. Category-level averages reveal broader trends, while individual outliers are examined to identify specific failure modes or exceptional responses. The combined dataset forms the basis of the results presented in Chapter 4. This approach strengthens the reliability of the evaluation by demonstrating not only how the model performs numerically but also why it behaves as observed.

4 Results

This chapter presents the empirical findings of the evaluation and explains what the results reveal about the capabilities and limitations of the CAI system in Pharmadata’s first-line support. The analysis is based on one hundred practical support queries drawn from historical tickets and evaluated using the criteria defined in Chapter 3.4.3. Appendix 2 provides the full dataset, including scores for clarity, accuracy, helpfulness, structure, terminology, escalation logic, and the number of prompts used in each interaction for each prompt. These metrics provide a systematic basis for examining how the LLMs behaves across different functional categories.

To keep the analysis closely aligned with the research problem, the chapter is organised so that each main section addresses one research question. Section 4.1 examines how accurately the CAI system can generate solutions to selected support requests in a real pharmacy IT context. Section 4.2 analyses how effectively the system distinguishes between cases that can be handled autonomously and those that require escalation to human support. Section 4.3 evaluates which types of first-line support requests are suitable for CAI-based automation and finally, Chapter 4.4 discusses the system-level structures and requirements needed to integrate a CAI-based support agent into support workflows. Together, these sections allow the chapter to respond to the research questions by linking quantitative performance results with task suitability, escalation behaviour, and system-level design implications for pharmacy IT support.

Before addressing each research question topic in detail, this section summarises the overall performance trends across the full evaluation set. Table 5 presents the average score for each functional category along with the overall average of all evaluation criteria. The scoring patterns reveal consistent strengths across several criteria and expose clear weaknesses in others. The six evaluation criteria defined in Chapter 3.5.1 are analysed both independently and in relation to the functional categories in which they occur. To structure the analysis and ensure transparency in interpreting the results, the averages are examined along two dimensions: (1) overall scores across the evaluation criteria and (2) category-level averages across the different support domains. This dual perspective enables the identification of general performance characteristics while also revealing category-specific strengths and limitations.

Across all 100 queries the overall performance of the CAI system reaches an average score of 3.64, representing the combined average of clarity, accuracy, helpfulness, structure, terminology and escalation logic scores in all functional categories. Rather than reflecting performance on any single task or

criterion, the score functions as an all-around indicator of the system’s general first-line support capability. Interpreted in context, on average responses were mostly correct, with many being partially or largely accurate. At the same time, the overall score conceals notable differences across evaluation criteria and functional categories. Performance varied most in tasks that required system-state awareness, coordination across multiple systems, or safety-critical judgment. The overall mean should therefore be interpreted as a baseline indicator of performance, which is examined in greater detail in the following sections through criterion-level and category-level analyses. Overall, the results suggest a moderate level of reliability

Table 5. Average Performance of the LLM Across Functional Support Categories. Note: This table presents the average evaluation scores of the LLM across functional support categories in Pharmadata’s service environment. The scores reflect performance on clarity, accuracy, helpfulness, structure, terminology, and escalation logic. The table also includes an overall average score for each category, showing how the model performed relative to different types of support tasks.

Functional category	Clarity	Accuracy	Helpfulness	Structure	Terminology	Escalation logic	Average of all evaluation criteria
Billing	4,26	3,89	3,47	4,32	4,32	4,26	4,09
Prescription Handling	4,45	3,82	3,27	4,36	4,64	3,91	4,08
Inventory Management	4,30	3,60	3,10	4,10	4,30	3,60	3,83
Device Issues	4,36	3,55	3,64	4,45	4,00	4,09	4,02
Miscellaneous	4,04	3,85	3,00	3,88	4,35	4,08	3,87
Access and Certificates	4,00	3,67	2,78	4,00	3,78	4,67	3,81
Reporting	4,33	4,33	4,00	4,50	4,17	4,67	4,33
Dose Dispensing	4,20	3,60	3,20	4,00	3,60	3,00	3,60
Network and Connectivity Issues	3,67	3,33	3,00	4,00	4,33	4,33	3,78
Average in all functional categories	4,20	3,78	3,25	4,16	4,23	4,09	3,64

4.1 CAI accuracy varies across categories

Accuracy is one of the most direct indicators of how well the CAI system can provide correct operational guidance. In this study, accuracy is defined as the degree to which the CAI system's response aligns with the reference solution derived from the historical training data. Each response was evaluated by comparing the model's suggested steps and conclusions against the correct resolution documented in the original support ticket. Accuracy therefore reflects alignment with established solutions rather than linguistic quality or user satisfaction. The model ability to give fully accurate guidance varied between categories. Figure 10 illustrates that accuracy reached an average score of 3.78 and is among the lowest-scoring evaluation criteria. This average falls below the values for clarity, terminology, structure, and escalation logic, making it the second weakest dimension of performance in the dataset. Although the score remains positive on the evaluation scale, it is clearly lower than the language-related criteria, which indicates that the model explains steps more consistently than it identifies the correct ones. This contrast already suggests that the clear formulations did not always translate into correct support outcomes.

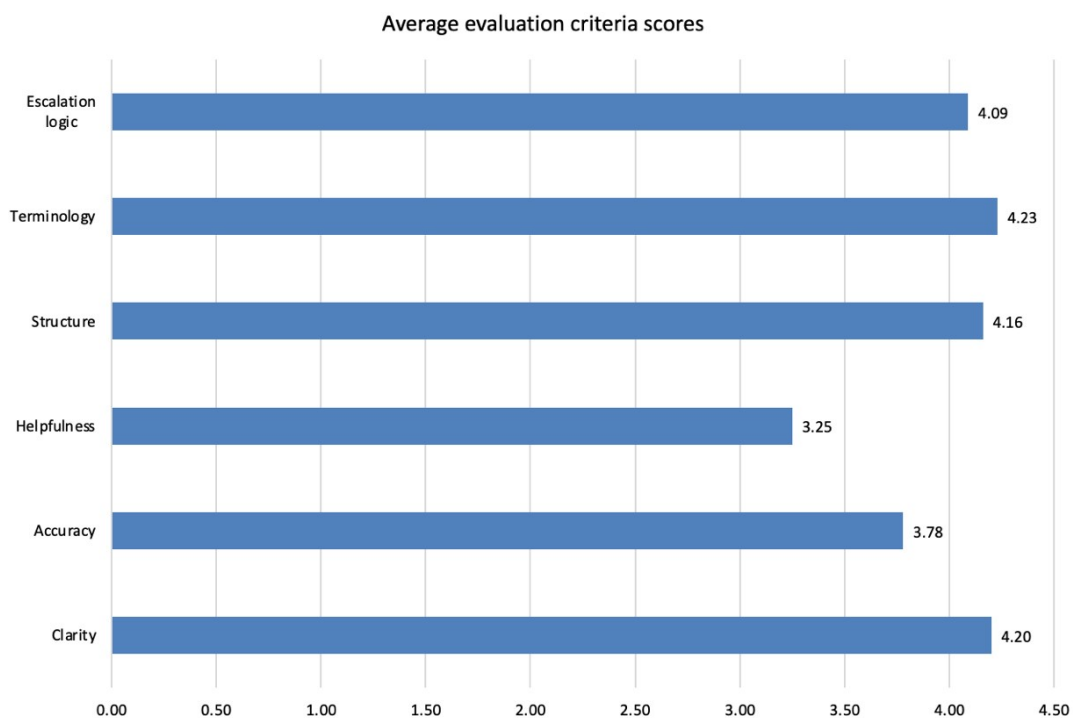


Figure 10. Average scores for each evaluation criterion across all functional categories. Note: Figure illustrates the average scores for each evaluation criterion across Pharmadata's functional support categories. Clarity, terminology, and structure received the highest scores, indicating strong performance in producing coherent and domain-appropriate

responses. Helpfulness scored noticeably lower, suggesting less consistency in providing actionable guidance. Overall, the results highlight both strengths and areas for improvement in the model's performance within Pharmadata's service environment.

While the average accuracy score provides a summary view of performance, the distribution of accuracy scores offers a more detailed picture of how often the CAI system produced fully correct guidance. Figure 11 presents the percentage distribution of accuracy scores across the 100 evaluated support requests. The results show that 62% of the requests received high accuracy scores (4 or 5), indicating strong alignment with the correct solutions documented in the historical support data. Of these, 30% achieved the highest possible accuracy score, reflecting cases where the model's response closely matched the reference solution without notable omissions or errors. At the same time, 27% of the requests received a mid-range accuracy score of 3, suggesting partially correct guidance. These captured the general direction of the solution but lacked a critical detail or verification steps. Lower accuracy scores were comparatively rare: 8% of requests scored 2, and only 3% scored 1, indicating that fully incorrect or misleading responses occurred infrequently. Overall, the distribution indicates moderate but uneven accuracy, with most responses aligning well with established solutions and a smaller share showing incomplete correctness.

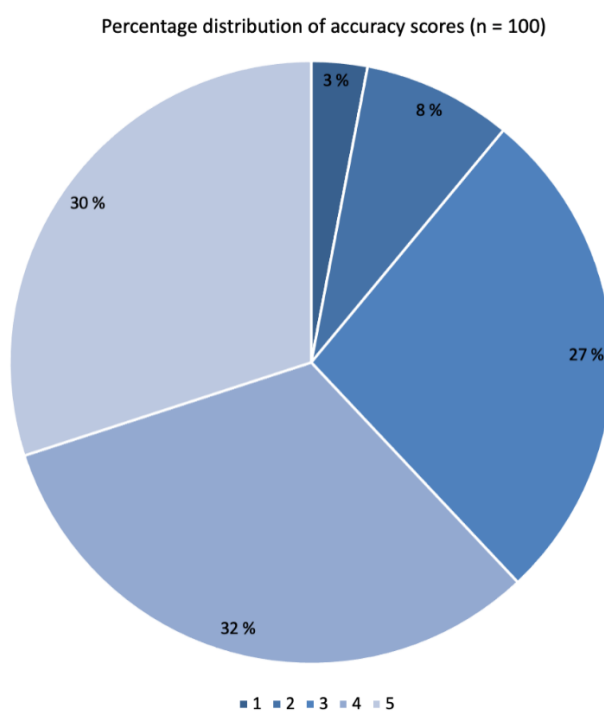


Figure 11. Percentage distribution of accuracy scores across evaluated support requests (n = 100). Note: The figure illustrates how accuracy scores are distributed across the evaluated support requests. Most

responses received high accuracy scores (4–5), indicating strong alignment with the correct solutions documented in the historical support data. A smaller proportion of responses received mid-range scores, reflecting partially correct guidance, while low accuracy scores were relatively rare.

Despite generally moderate accuracy overall, performance varies significantly across functional categories. Figure 12 shows that the strongest accuracy appears in reporting, which reaches 4.33 and stands out as the clearest high-performing category. Billing (3.89) and prescription handling (3.82) also perform well. These categories share stable, well-defined workflows that rarely require context-specific adjustments or real-time inspection of system data. Because these processes remain consistent across pharmacies, the CAI system can map user queries to familiar procedural patterns and reproduce the steps more reliably. In contrast the lowest accuracy is most evident in categories that require system validation or domain-specific checks. Network and connectivity issues score 3.33 forming the weakest category regarding accuracy. These tasks often depend on diagnosing hardware conditions, evaluating network status, or interpreting error codes. All information that the CAI model might not have access to. As a result, responses tended to lean toward general troubleshooting steps, which often fell short of the required precision. Device issues show a similar limitation scoring 3.55, the second lowest accuracy value in the dataset. As a result, responses tended to lean toward general troubleshooting steps, which often fell short of the required precision. Accuracy also remains limited in dose dispensing and inventory management, both scoring 3.60. Inventory-related tasks often require checking stock configurations or product settings, while dose dispensing queries involve medication-specific or clinically sensitive rules. These tasks demand precise verification steps that the CAI cannot perform, which leads to more frequent inaccuracies.

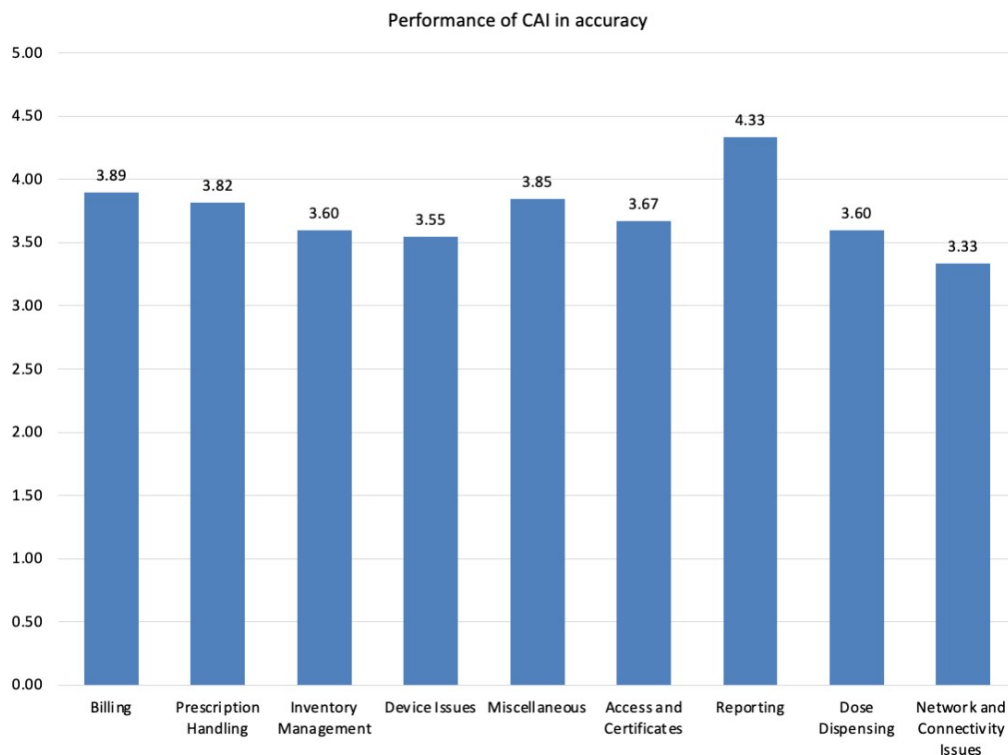


Figure 12. Comparison of average accuracy across functional categories. Note: This figure compares the average accuracy of the LLM across Pharmadata’s functional support categories. Reporting achieved the highest accuracy, while categories such as device issues and network or connectivity issues scored lower. The results highlight variation in how reliably the model handles different types of support tasks.

Patterns visible in Appendix 2 reinforce these category-level differences. Tasks that rely on system-state inspection, backend validation, or context-specific configuration steps consistently produced lower accuracy because the CAI system cannot access the underlying information needed to confirm its reasoning. These tasks also tend to occur less frequently in historical support data, which limits the model’s exposure to stable solution patterns. When the model has few similar cases to draw from, it tends to fall back on broad troubleshooting patterns that do not fully resolve the problem. In contrast, higher accuracy appears in categories grounded in stable, well-documented procedures. These tasks are less dependent on real-time system information and more aligned with the model’s available training patterns.

These accuracy patterns show that the CAI system performs moderately well in domains grounded in stable, predictable procedures. It performs less reliably in tasks that require real-time diagnostic insight or specialised domain knowledge. These differences form the empirical basis for assessing which

task types are suitable for automation and which require human oversight. Section 4.3 examines these implications in detail.

4.2 Escalation behaviour and boundary recognition

A key requirement for safe CAI-based automation is the system's ability to recognise when it lacks sufficient information or when a query involves safety-critical, configuration-specific, or regulatory-sensitive tasks. Escalation behaviour therefore forms a direct indicator of whether the CAI can distinguish between tasks that may be automated and those that must be transferred to human support. The evaluation shows a generally strong performance in escalation logic, but the results also reveal category-specific inconsistencies that limit autonomous task handling in several areas.

4.2.1 Performance in escalation logic

Escalation logic performs strongly across categories. Figure 10 reveals that the average score of 4.09 places it among the highest evaluation criteria in the dataset. The high score indicates that the model often identifies its limitations correctly. It also shows that the system does not rely on overconfident reasoning in cases where it lacks sufficient context to provide a reliable answer. This behaviour is essential in first-line support, where technical uncertainty, regulatory constraints, and patient-safety considerations require strict boundaries between automated and human-controlled tasks (Li et al., 2024).

As shown in Figure 13, escalation scores remain consistently strong across functional categories although the values vary a bit. The highest escalation values are found in reporting and access and certificates, both reaching a score of 4.67. These results show that the CAI reliably identified tasks requiring human verification, particularly those involving identity management, permission changes or certificate validity. Network and connectivity issues also show a high escalation rate (4.33), reflecting that the CAI consistently recognised when diagnostic steps required inspection of logs or hardware states that it could not access. In the opposite end of the range, dose dispensing produced the lowest escalation score (3.00) and inventory management (3.60). While these are lower than the highest-scoring categories, they still represent acceptable performance within the evaluation scale. When looking more deeply into Appendix 2 it is visible that dose dispensing had the most inconsistent escalation behaviour. In some medication-related cases the CAI attempted to solve queries that should have been transferred to a human agent, while other, similar tasks were escalated immediately. Across categories, unclear queries and tasks requiring checks of logs, device conditions or certificate validity were prone to escalation errors. This indicates the most

consistent escalation in different categories. The spread between the lowest and highest escalation scores shows that escalation behaviour is not uniform across all task types. Escalation is an essential feature in technical and safety-critical environments because it acts as a boundary between tasks that can be automated and tasks that require an interpretation of system states, regulatory constraints, or patient-related information (Li et al., 2024).

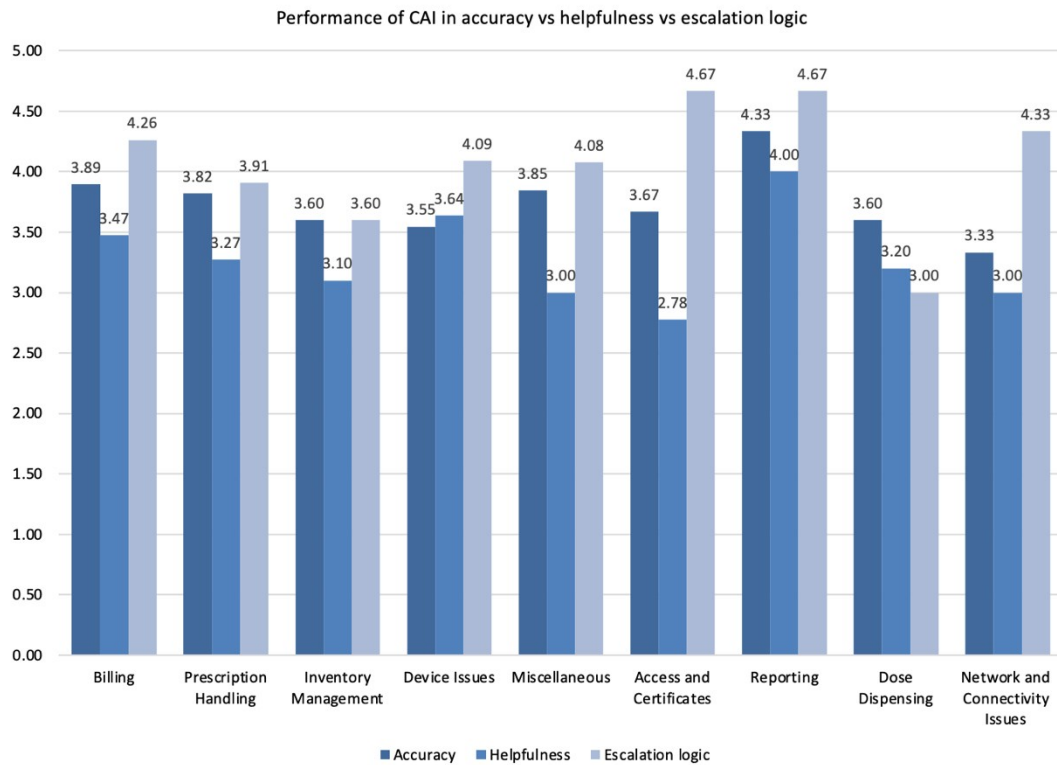


Figure 13. Comparison of average accuracy, helpfulness, and escalation logic across functional support categories. Note: This figure compares the LLM’s accuracy, helpfulness, and escalation logic across functional support categories. Reporting shows the strongest overall performance, while areas such as access and certificates, dose dispensing, and network or connectivity issues perform lower. The comparison illustrates how the model’s strengths and weaknesses vary based on task categories.

Correct escalation rates across functional categories provide a complimentary view for patterns based on escalation scores. Figure 14 visualises how often the CAI escalated tasks correctly in each category. Reporting and network and connectivity issues achieved a 100 percent correct escalation rate. In these categories, the system systematically recognised when the task required human verification or system-state inspection. Access and Certificates also show a high escalation rate at 89 percent. The results for billing, device issues, inventory management and miscellaneous tasks fall between 80 and 89 percent. These tasks include both routine queries that can be answered

safely and more complex cases that require environment-specific verification. The model escalated most cases correctly, which aligns with the operational need to avoid accidental misconfigurations or unauthorised changes.

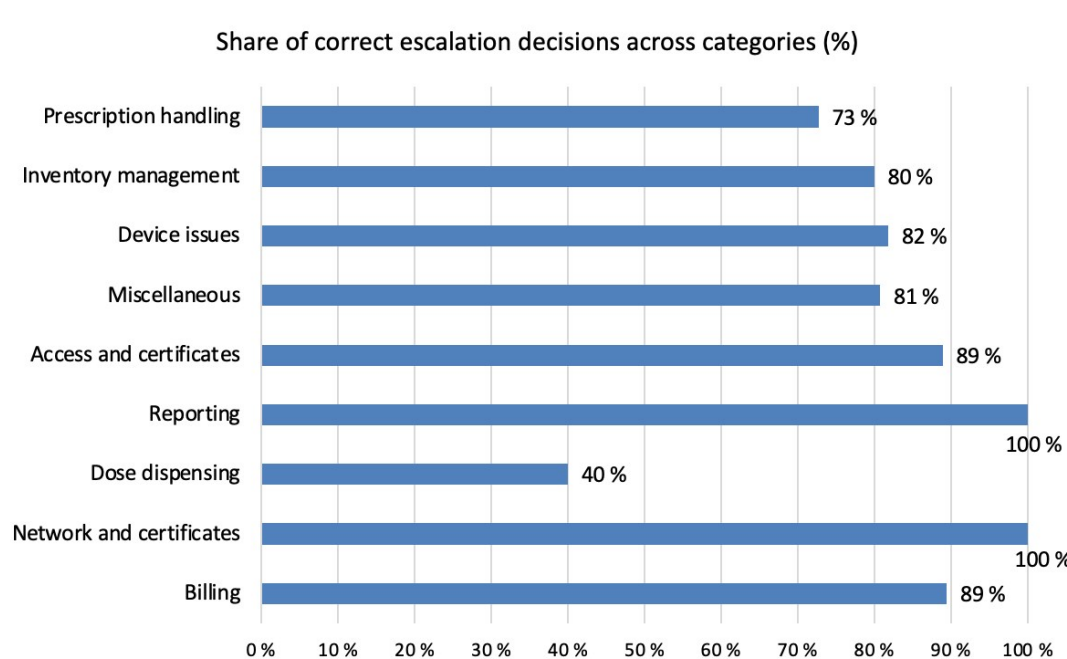


Figure 14. Percentage of correct escalation decisions across functional support categories. Note: This figure shows the share of test queries in each functional category for which the LLM selected the correct escalation decision. Reporting, network and connectivity issues, and billing reached 100 percent correct escalation, while dose dispensing showed the lowest share. The percentages illustrate how consistently the model recognised escalation boundaries across different types of support tasks.

By contrast, dose dispensing stands out with a correct escalation rate of only 40 percent. This category includes medication-related workflows that are governed by strict safety rules. These tasks may require checking dosage constraints, product-specific information or clinical logic that the CAI cannot verify. The low escalation rate indicates that the model did not consistently identify the need for human involvement, which poses operational and clinical risks.

4.2.2 Interaction between escalation, accuracy and helpfulness

Helpfulness is the lowest scoring criteria in the evaluation, with an average of 3.25 across all functional categories (Figure 10). As shown in chapter 4.1 the lowest accuracy scores appeared in domains such as dose dispensing, network and connectivity issues, and inventory management. A similar pattern emerges in the helpfulness results. The weakest values appear in access and certificates (2.78), network and connectivity issues (3.00), miscellaneous

(3.00) categories. These largely overlap with the categories where accuracy declined. Low helpfulness reflects either that the inaccurate answer did not assist the user or that the model did not attempt a full solution because it escalated the query instead.

A closer look at Appendix 2 reveals that several individual queries with low helpfulness scores (2-3) appear together with high escalation logic scores (4-5). This pattern appears most clearly in categories such as access and certificates and dose dispensing, where correct handling depends on information the model cannot access. These include certificate validity, permission settings or medication-related checks. In these situations, a low helpfulness score does not necessarily signal poor performance. Instead, the combination of low helpfulness and high escalation indicates that the model avoided offering incomplete guidance and correctly transferred the case to human support. In other words, the model functioned as intended. This is particularly evident in cases where accuracy remains relatively high, but helpfulness is low because the model did not attempt to complete the workflow. Here the low helpfulness reflects the absence of a final solution rather than an incorrect answer. The CAI recognised that it could not safely continue the task and escalated instead. This behaviour aligns with the escalation criteria defined in Chapter 3.5.1 and shows that escalation logic can operate as a compensating mechanism when the model reaches the limits of its capability.

Across these categories, patterns in Appendix 2 show that many of the weaker accuracy and helpfulness scores occur in tasks that are safety-critical or occur infrequently in the original support ticket dataset. As a result, the model encounters fewer consistent examples from which to learn the correct sequence of actions. This lack of exposure is visible in the results: when no close precedent exists in the training material, the CAI tends to rely on broad patterns or general troubleshooting logic, rather than producing precise context-specific steps. This behaviour appears most clearly in access and certificates, network and connectivity issues, and dose dispensing. In these categories many queries depend on procedures that change over time, require detailed verification, or come up only in rare situations. These characteristics help explain why accuracy and helpfulness lag other criteria in Figure 10 and why Appendix 2 includes several cases where the CAI's guidance remained generic even though the task required specific domain knowledge. The broader implications of this pattern for task suitability and safe automation are discussed in Section 4.3.

4.3 Suitability of CAI for service desk support in pharmacy IT

The patterns observed in Chapters 4.1 and 4.2 establish the foundation for assessing which types of first-line support tasks can be automated safely and effectively using CAI. While accuracy and escalation behaviour are central to this assessment, they do not on their own determine suitability. A complete evaluation also requires examining how the CAI performs in several additional dimensions that directly influence its operational feasibility in a regulated healthcare IT environment. These include the model’s language-related capabilities, its performance across different functional task categories, and the efficiency with which it resolves queries.

Together, these results capture the broader set of factors that shape whether CAI can operate reliably within pharmacy IT, where tasks vary in their dependence on system-state information, safety-critical reasoning, and domain-specific terminology (Wong et al., 2025). This section therefore presents the remaining empirical findings needed to evaluate suitability by analysing language performance, category-level outcomes, and prompt efficiency across the full evaluation dataset.

4.3.1 Language performance as a precondition for suitability

To assess which support tasks are suitable for CAI-based automation, it is necessary to examine how reliably the agent communicates technical instructions. This section therefore reports the results for the language related criteria: clarity, structure and terminology across functional categories. The strongest performance occurs in these categories. As shown in Figure 10, the highest averages appear in the language-related criteria: clarity (4.20), terminology (4.23), and structure (4.16). These results show that the LLM model can express instructions clearly, maintain a logical structure and use technical terminology correctly in the majority of support scenarios. Figure 15 further illustrates this pattern by showing that clarity, structure and terminology remain consistently strong across most functional categories. This stability suggests that the model’s language generation capability is largely unaffected by category differences.

A few exceptions fall below 4.0 which still reflects strong performance but points to slight variability across categories. These exceptions are still meaningful since they indicate the boundary between routine tasks where the model communicates reliably and tasks where language performance becomes less stable. The terminology scores in access and certificates (3.78) and dose dispensing (3.60) are the clearest examples. Both categories rely on domain-specific vocabulary that must be used with precision. Access tasks require accurate references to certificate types, permission levels and integration components. On the other hand, dose dispensing depends on clinical terminology, dosage rules and product naming. In both cases, the model

occasionally generalises or misuses these terms. Network and connectivity issues show a different pattern with a clarity score of 3.67. These problems often require information that the model cannot access, such as error logs, device states or configuration details. Without this context, it gives general troubleshooting steps. These answers may be correct in principle, but they lack the detail needed for accurate diagnosis. The lower clarity score reflects this limitation and shows that the model might struggle when essential system information is unavailable.

Overall, these results show that the model’s language performance is strong but not consistent across all domains. High scores in clarity, structure and terminology demonstrate that the model can produce professional and well-formed responses, which is a basic requirement for CAI automation. But it must be taken into account that strong language does not guarantee operational usefulness. The under 4.0 scores reveal that performance declines when tasks rely on real-time system states, clinical or regulatory precision, loosely structured queries or specialised terminology. These variations are relevant for assessing task suitability and are examined further in Chapter 4.3.3.

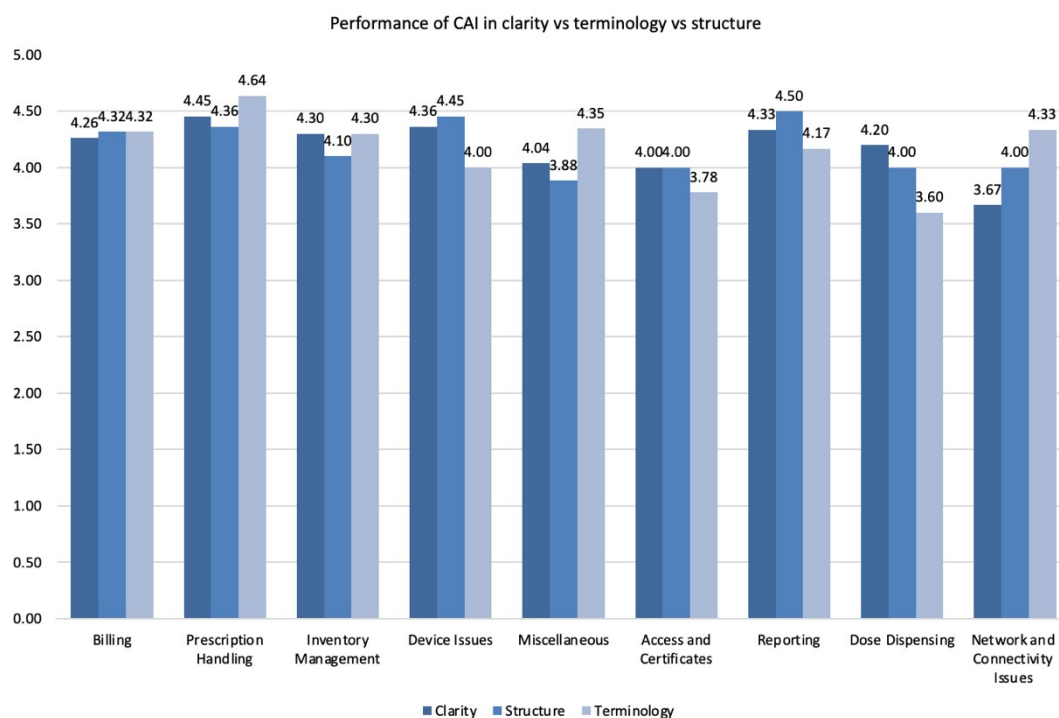


Figure 15. Performance of CAI model in clarity, structure, and terminology across functional categories. Note: This figure compares the LLM’s clarity, structure, and terminology scores across Pharmadata’s functional support categories. Reporting and prescription handling show the highest consistency across all three criteria, while dose dispensing and network or

connectivity issues display more variation. The results highlight differences in how well the model communicates, organises information, and applies domain-specific terminology in different task areas.

4.3.2 Functional category-level performance differences

To identify where CAI is most suitable for first-line support, this subsection reports the average overall performance score in each functional category and the corresponding prompt counts. Figure 16 visualises that the strongest performance appears in categories with well-defined procedures. Reporting achieves the highest average score (4.33), reflecting the fixed and repeatable nature of the workflow. Device issues also perform well (4.02), consistent with the model's ability to follow standard troubleshooting steps. Billing (4.09) and prescription handling (4.08) similarly show strong performance benefitting from stable terminology and structured processes that align with the CAI system's language level skills.

Performance drops slightly in more complex cases where support work requires reading logs, checking integrations, or applying clinical rules such as dose dispensing (3.60) and network and connectivity issues (3.78). Dose dispensing tasks often involve clinical or safety-critical decisions, and errors can create real risks. The model does not always detect that the case is sensitive and should not be handled by a CAI agent. As visible from Appendix 2 these individual issues in accuracy and escalation logic results in decreasing average score for CAI performance in dose dispensing category. Similarly with network and connectivity issues that at times depend on diagnostics the model cannot perform, such as checking hardware states or examining logs. Human support is often needed to confirm the cause of the failure and therefore the helpfulness of CAI in this category is limited. This distinction is central for understanding which task types are suitable for automation and where human judgement remains essential.

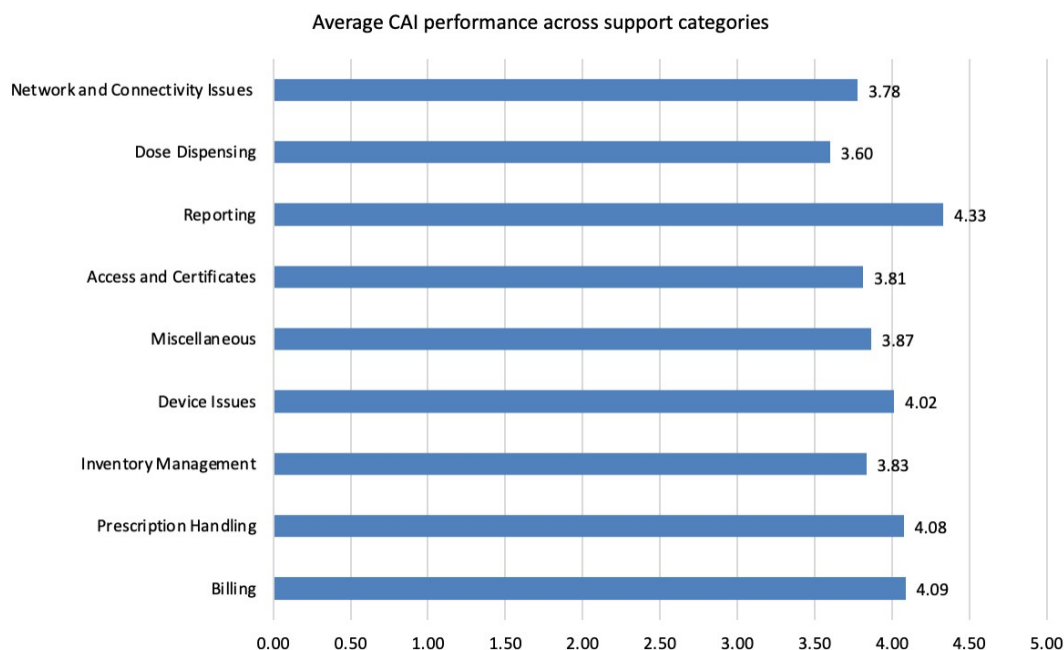


Figure 16. Average CAI performance across functional support categories. Note: This figure presents the overall average performance scores of the LLM across Pharmadata’s functional support categories. The results summarise how well the model performed across all evaluation criteria for each task category.

As indicated in Figure 16 mid-range results appear in inventory management (3.83), the miscellaneous category (3.87), and access and certificate tasks (3.81). These categories mix simple queries with tasks that depend on system states or user-specific information. The model explains general procedures clearly, yet it struggles when the correct answer depends on information it cannot verify or access. Access-related tasks highlight this point most clearly, since many of these cases require backend checks and actions.

The number of prompts used per CAI discussion offers another view of how efficiently the model works in practical support situations. Figure 17 visualises this pattern by presenting the overall average prompt count (1.78) in a brighter colour at the top of the chart, with each functional category average prompt count displayed beneath it. This layout enables direct comparison between the global average and the category-specific values that form it. The overall average of 1.78 prompts indicates that most queries were resolved within one or two exchanges. This suggests that the model can often produce a complete and coherent answer without extended back-and-forth interaction. Short prompt sequences are particularly common in categories with stable and well-understood workflows such as access and certificates (3.81), where the required steps are highly standardised. A slightly higher prompt count appears in categories such as reporting (2.50) and device issues (2.09)

where the queries are often more complex and might not have direct links to instructions. These tasks require details that the model cannot infer, including error messages, device states, or permission levels. When this information is missing the model at times asks for clarification before it can provide a more reliable answer. The prompt count therefore helps distinguish between tasks the model can solve directly and tasks that require more user input.

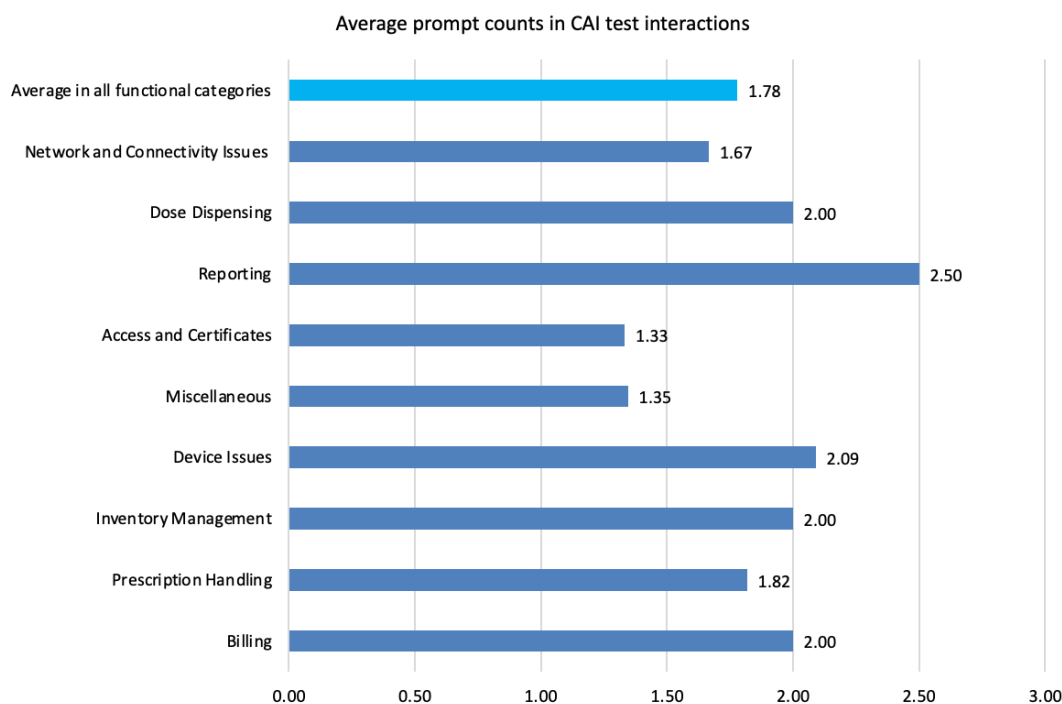


Figure 17. Average and category-level prompt counts in CAI test interactions. Note: This figure shows the average number of prompts the model required to respond to queries in each functional support category, illustrating how interaction length varies by task type. The light bar represents the overall average across all categories, while the darker bars show category-specific averages. Reporting and device issues required the most prompts, whereas access and certificates and miscellaneous tasks required fewer.

Taken together, the category-level performance scores and prompt-count patterns show clear variation in how efficiently the CAI handles different types of support requests. Higher scores and shorter prompt sequences appear in domains with well-defined, repeatable workflows, while lower scores and longer interactions occur in tasks that depend on system-state checks, backend information, or specialised procedural knowledge. These distinctions are used in Chapter 4.3.3 to delineate task types that are suitable for automation.

4.3.3 Tasks suitable for CAI automation

The evaluation results show clear variation in the CAI system's performance across different functional categories, allowing an assessment of which task types demonstrate higher or lower suitability for automation according to the measured criteria. Categories with consistently high scores across accuracy, clarity, structure, terminology, helpfulness and escalation, combined with low prompt counts, form the strongest group.

A key enabler of this suitability for automation is the alignment between CAI-generated guidance and Pharmadata's Neuvo and Omaneuvo documentation. Many of the procedures in these categories are thoroughly documented in these systems, and the CAI frequently directs users to the relevant links. The results in Chapters 4.1, 4.2, 4.3.1 and 4.3.2 reinforce the importance of thorough documentation: in these task types, clarity, structure and terminology remain high, accuracy is strong, and prompt counts are low. Together, these patterns indicate that the CAI recognises user intent efficiently and can map queries to familiar solution pathways without extended clarification.

These findings indicate that several types of support tasks are well suited for CAI-based automation. Reporting and billing are the clearest examples. Figure 16 shows that these categories reach the highest overall averages (4.33 for reporting and 4.09 for billing), combining strong clarity, structure and terminology with solid accuracy and helpfulness. In practice this means that the agent not only explains the steps correctly but usually selects the right sequence of actions when users need to run standard reports or resolve common billing issues. Figure 17 supports this by showing that these tasks are often completed within one or two prompts. Many of these queries concern initial diagnostic steps represented in the historical ticket dataset and do not require long clarification before providing a workable answer. Routine prescription handling also fits this group. Many of the evaluated queries concerned navigation in Pharmadata's systems or explanation of existing processes rather than clinical decision-making. In these cases, the model benefited from high terminology scores and from the fact that the underlying workflows are documented in Neuvo. When users are guided either directly through the steps or directed to the relevant Neuvo article, they receive instructions that are both familiar and consistent with existing support practices.

Basic device troubleshooting shows similar potential. The first diagnostic steps, such as checking power, connections or basic settings, form a limited and stable set of actions. Figure 16 reports an overall average of 4.02 in this category, and Figure 17 shows that most device-related queries are solved in short prompt sequences. This pattern suggests that the CAI can reliably

handle the initial phase of troubleshooting, while more complex hardware-specific issues continue to be escalated to human support. Inventory management tasks occupy a slightly more mixed position, yet the figures indicate that a substantial subset is suitable for automation. The overall average of 3.83 remains clearly positive, and language-related criteria perform well. Tasks that involve explaining ordering workflows, locating inventory functions in pd3 or interpreting product lists align with these strengths. The evaluation also shows that problems arise mainly when the correct answer depends on real-time stock levels or configuration data that the model cannot see. For this reason, explanation and navigation tasks in inventory management can be automated, whereas verification steps and configuration changes should remain with human agents.

A specific nuance appears within access and certificates. Although this category involves many tasks that depend on backend information and are therefore unsuitable for automation, Appendix 2 revealed that some queries fit CAI's capabilities. These include guiding users through password changes, supporting basic account navigation, locating certificate-related information and retrieving Neuvo documentation links. These steps are well defined and do not require system-state checks.

In contrast, several categories include tasks that cannot be handled autonomously by a CAI system. These tasks rely on information that the model cannot retrieve, such as permission levels, certificate validity, backend configurations, system logs, error codes or network states. Because the model has no visibility into these system elements, it cannot verify its assumptions, and its output remains speculative. The access and certificates category illustrates this boundary clearly. Tasks that involve validating permissions, verifying certificate status or adjusting backend configurations cannot be automated because the CAI cannot observe the system environment and therefore cannot confirm whether its assumptions are correct. Human oversight is required for any task that alters system access, interacts with identity and certificate infrastructure or has implications for security.

Network and connectivity issues show similar limits. Effective troubleshooting often depends on reading logs, verifying network states or inspecting device-specific information. None of this is visible to the model. The CAI responds with general advice that may be correct in principle but lacks the specificity needed for accurate diagnosis. These tasks also displayed higher prompt counts in the evaluation, indicating that the model repeatedly asks for missing information that it cannot use to reach a reliable conclusion. Dose dispensing forms the most sensitive category. These tasks require understanding of clinical rules, dosage constraints and product-specific properties. The performance evaluation showed lower accuracy and less consistent

escalation in this domain. Errors in medication-related workflows have direct safety implications. Even when terminology is correct, the model does not always identify that a task is clinically sensitive and should not be handled autonomously. Medication-related decision pathways therefore fall outside the scope of safe automation.

Across these task types, the evaluation indicates a consistent requirement for human oversight whenever task completion depends on hidden system information, backend operations, regulatory considerations or clinical constraints. In such cases, the CAI may still be valuable for explanation, initial triage, or retrieving relevant Neuvo documentation, but responsibility for final action must remain with human staff who can verify system states and assess risks.

4.3.4 Regulatory and operational constraints shaping suitability

Pharmacy IT support operates within a regulatory environment that places strict requirements on the handling of data, user authentication, and medication-related information (Wong et al., 2025). Several functional categories evaluated involve tasks with strict compliance demands, such as access rights, certificate validation, and medication-related steps. These tasks must follow GDPR principles, which require data minimisation, anonymisation when possible, and clear rules for how information is processed (European Commission, 2025). Governance frameworks such as AI TRiSM add further expectations by emphasising controlled decision boundaries and protection against unsafe outputs (Habbal et al., 2024). The broader European legislative environment reinforces these expectations (European Commission, 2025). The GDPR regulates the handling of sensitive health data, and the AI Act places strict requirements on high-risk systems in medical and pharmaceutical contexts, including transparency, human oversight, and reliable data use (European Commission, 2025).

As highlighted in Section 4.2.1, escalation scores are strongest in categories that require strict control of identity-related workflows or system-state verification. These tasks often require verification of system states, inspection of logs, or checks against regulatory requirements. Because the CAI has no real-time visibility into these elements, it cannot reliably confirm whether its proposed steps are correct. This makes fully automated handling inappropriate. By contrast, tasks in reporting, billing, and device troubleshooting rely on well-documented procedures and do not require interaction with protected data or sensitive configuration elements. These characteristics make them more compatible with regulatory expectations and therefore more suitable for CAI-based automation.

The escalation patterns presented earlier reflect these regulatory structures. Figure 14 shows that the CAI escalated correctly most often in categories where regulatory constraints are strongest, such as tasks involving identity verification, certificate handling and network-level diagnostics. In these areas, GDPR and the AI Act require that decisions remain traceable and verifiable by a human operator. The CAI's behaviour aligns with these requirements: when a task depended on protected data or system-state information that the model cannot access, it frequently escalated rather than attempting a speculative solution. This pattern suggests that the model respected the operational boundaries shaped by regulation.

By contrast, Figure 14 also highlights the one domain where escalation was inconsistent: dose dispensing. Medication-related workflows involve dosage rules, product-specific safety considerations and clinical judgement. These tasks fall under the strictest regulatory controls, and incorrect handling carries direct patient-safety risks. This confirms that medication-related tasks fall outside the safe scope of CAI automation and require continuous human oversight. For routine, well-documented tasks such as reporting, billing and standard troubleshooting, the regulatory demands are lower because these workflows do not involve sensitive data or backend configuration checks. These areas also showed stable escalation behaviour in Figure 14. Their compatibility with CAI automation therefore rests not only on procedural stability, but also on the fact that performing these tasks does not violate regulatory boundaries.

4.3.5 Patterns relevant to reliability

Trust in a CAI system depends on the degree to which its behaviour is consistent, predictable and aligned with the boundaries of safe task execution. The evaluation shows that the model's performance supports trust in several ways. The model's strong clarity, structure and terminology scores across the dataset visualized in Figure 15 create a stable foundation for user confidence in routine interactions, where instructions must be readable and aligned with the terminology used in pharmacy IT systems. These language related strengths mean that users can rely on the system to communicate technical content clearly, which is necessary for safe and efficient support work.

However, trust does not arise from language quality alone. Users must also be confident that the system recognises when a task exceeds its competence. Here the escalation patterns observed in Chapter 4.1 and Chapter 4.2.2 provide important insights. In many cases, particularly in categories such as access and certificates, reporting and network and connectivity issues, the model escalated correctly and consistently. These cases demonstrate that the system can identify tasks that require human verification, system-state

checks or regulatory sensitivity. High correct escalation percentages in these domains (Figure 14) indicate responsible behaviour that supports safe user reliance. At the same time, the performance evaluation also revealed a small set of cases where trust could be compromised. Almost all functional categories contain a small number of cases where low accuracy, low helpfulness, and low escalation logic occur simultaneously. Appendix 2 shows that such instances are not frequent, yet they appear across several domains, including prescription handling, device issues, miscellaneous tasks, dose dispensing, access and certificates, and network and connectivity issues. Although these cases represent individual errors rather than systematic patterns, they are analytically significant because they illustrate situations where the model produced incomplete or incorrect guidance and did not recognise that escalation was necessary. This combination is problematic because inaccurate steps not only fail to resolve the user's problem but also disrupt the need for human involvement. Although these cases remain exceptions rather than systemic patterns, they illustrate that users cannot rely on the CAI without maintaining critical oversight.

Overall, the results show that CAI agent is a valuable tool for a substantial portion of Pharmadata's first-line support workload. Its strongest contributions emerge in domains where workflows are standardised, terminology is stable, and actions do not depend on real-time system information. Reporting, billing, prescription handling, device troubleshooting, and several inventory-related tasks represent the most suitable targets for automation. In these categories, the CAI provides clear, well-structured responses, aligns closely with Neuvo documentation, and resolves queries efficiently, often within one or two prompts. Tasks requiring backend verification, configuration checks, authentication handling, or medication-related reasoning must remain with human agents, as the system cannot access the contextual information these workflows require.

4.4 System-level constraints affecting CAI performance

The findings presented in Appendix 2 highlight several system-level constraints that affect CAI performance in a pharmacy IT support context. These principles arise from the model's observed strengths and weaknesses visible across Figures 10-17 and reflect the operational, regulatory, and user-experience requirements of a pharmacy IT support environment. Rather than proposing detailed solutions, this section summarises the main technical and data-related factors that shape what a CAI system can realistically do as a first-line support agent. The practical design choices based on these constraints are discussed further in Chapter 5.

A key system constraint is the CAI agent's lack of access to system states, configuration settings, and live error logs. The model responds only with the information provided in the prompt and the patterns it has learned from historical support tickets. It cannot check whether a certificate has expired, whether a user has the correct permissions, or whether a device is offline. These gaps create systematic blind spots that appear across several functional categories in Appendix 2. The evaluation shows that these blind spots lead to incomplete or incorrect answers in tasks that depend on real-time system information. The pattern is especially clear in access rights, certificate validation, network diagnostics, and dose dispensing. These tasks normally require manual work from service desk staff, because correct handling depends on checking system output, opening configuration menus, or reviewing device logs. The CAI agent cannot see any of these elements. It can only guess based on the wording of the user's question and the types of cases it has seen before. As a result, the model may produce instructions that sound correct but do not fit the actual situation. These errors are not caused by weaknesses in vocabulary or sentence structure. They arise entirely from missing operational context. The results therefore indicate that tasks which depend on system checks, user permissions, certificate states, device conditions, or medication-related decision steps are systematically more vulnerable to error.

A second limitation concerns how the escalation criteria operated in practice. As presented in Figures 13 and 14, the overall escalation score was high and suggested reliable behaviour. However, the detailed results in Appendix 2 show that individual decisions were not always consistent. In some cases, the CAI escalated appropriately, while in others similar tasks did not trigger escalation even when the model showed uncertainty or produced only partially correct guidance. The evaluation also includes examples where the system escalated tasks it could have handled correctly based on the available information. These variations do not indicate a systematic error, but they show that the boundary between tasks suitable for autonomous handling and those requiring human involvement is not yet fully stable. This variability matters for system design. Escalation is the main safeguard when the agent has no access to real-time system-state information, and inconsistent use of this safeguard reduces the reliability of the hybrid support model. The CAI performance evaluation therefore identifies escalation behaviour as an area with clear operational constraints and as a feature that will require more robust mechanisms. Chapter 5 further examines the implications of these escalation patterns for future system logic.

5 Conclusion and discussion

5.1 Discussion of key findings

Based on the research findings, the conclusions of this thesis are drawn by answering each of the research questions. The results show that CAI can support first line pharmacy IT work, but its benefits appear only in specific conditions. The main conclusions are:

1. How accurately can a CAI system generate solutions to selected support requests in a real-world pharmacy IT context?

The Copilot Studio CAI agent can answer routine support cases with good accuracy, but it cannot handle tasks that rely on backend actions such as performing updates or access information that only a human support agent can retrieve. Its performance also declines in multi-step scenarios and in topics that involve medication safety.

2. How effectively can a CAI system distinguish between cases suitable for automation and cases requiring human escalation?

The CAI can often recognise when a task requires escalation, particularly in categories involving identity management, certificates, or network diagnostics. However, escalation behaviour is inconsistent at the level of individual tasks: similar tasks do not always receive the same escalation response. The system therefore shows partial but not fully reliable ability to distinguish between tasks suitable for autonomous handling and those requiring human involvement.

3. What types of first line in pharmacy IT support requests are suitable for automation using CAI?

Tasks that are low risk, follow repeatable and well-documented workflows, and do not require backend verification are most suitable for automation. These include reporting procedures, routine billing issues, navigation tasks in prescription workflows, basic device troubleshooting, and a subset of inventory-related queries. Tasks dependent on system-state checks, permissions, certificate validation or medication-related logic are not suitable for automation.

4. What system-level structures and requirements are necessary to integrate a CAI-based support agent into existing pharmacy IT workflows?

Integrating a CAI based support agent into pharmacy IT workflows is feasible when the organisation has a reliable knowledge base, clear hybrid processes,

and strong compliance with EU data and AI regulations. Successful adoption also requires planned change management and often support from an external technical partner, since the CAI must work securely within the existing IT environment while leaving higher risk decisions to human agents.

The following sections discuss each research question in more detail.

Research question 1. How accurately can a CAI system generate solutions to selected support requests in a real-world pharmacy IT context?

The results of this study show that the CAI model can answer many pharmacy IT support requests with a high level of linguistic quality, but its factual accuracy varies significantly across task types. The language quality consistently scored above four out of five, which indicates that the model can communicate in a professional and domain-appropriate way. It recognised common workflows in pd3 and Omapd, explained routine tasks clearly, and produced answers that matched the tone of first-line support. These results confirm that the model can operate effectively when it works within stable, well-understood procedures.

The model's accuracy was highest in cases where the prompt described the required steps explicitly or where the historical ticket data contained many similar examples. In such situations, the CAI reproduced the correct sequence of actions and provided responses that would have resolved the user's issue. These patterns were most evident in categories such as billing, reporting, device troubleshooting, and routine navigation tasks, where procedures rarely change and support agents typically follow fixed step-by-step workflows. Accuracy declined in tasks where the input lacked detail, where processes varied between pharmacies, or where the model had limited exposure to comparable historical cases. In these situations, the CAI often produced responses that were only partially correct. Errors included missing essential steps, inferring details incorrectly, or recommending procedures that were not consistent with the solutions used by Pharmadata's support team. Although these responses were fluent and well structured, they would not have fully addressed the underlying issue. The distribution of accuracy scores provides additional insight into the nature of these limitations. Most inaccuracies were partial rather than severe, with fully incorrect responses occurring infrequently. This indicates that the CAI typically captures the general logic of the solution, even when it fails to reproduce all required steps or verification checks. From a first-line support perspective, this pattern is significant: partially correct guidance may still reduce support workload or guide users toward resolution when combined with appropriate escalation mechanisms,

whereas consistently incorrect responses would pose a higher operational and safety risk.

The lowest accuracy appeared in tasks that rely on system-state information or real-time verification. Without access to logs, configuration settings, permissions, or device status, the model occasionally generated steps that sounded plausible but did not match the technical requirements of the situation. Functional categories that depend on backend validation showed the clearest limitations. Medication-related tasks also resulted in lower accuracy because the model could not reliably reproduce the detailed rules and product-specific constraints required for these cases.

Overall, the findings indicate that the CAI system achieves high accuracy in routine first-line support tasks with stable, repeatable workflows but shows notable accuracy variation in cases requiring dynamic system information or specialised domain knowledge. This establishes a clear boundary between task types the CAI can handle consistently and those where its factual correctness declines. This boundary is central for interpreting the model's role in real-world support work.

Research question 2. How effectively can a CAI system distinguish between cases suitable for automation and cases requiring human escalation?

The findings of this study show that the CAI system can identify the need for escalation in many routine support scenarios, but its reliability varies across task types. In the evaluation, each support query was assessed for its accuracy, clarity and whether the CAI attempted to solve the problem or escalated it to a human agent. Across the dataset, the CAI escalated reliably in tasks that clearly depended on information it could not access. These included identity verification, certificate handling and several types of network and connectivity diagnostics. In these cases, the model often recognised that backend checks or system-state information were required and indicated that the task should be handled by human support.

However, the results also show that this behaviour was not consistent across all tasks. Individual queries within the same category sometimes produced different outcomes, even when the underlying conditions were similar. This inconsistency was most visible in medication-related tasks, such as dose dispensing. These tasks required detailed domain knowledge or verification steps that the CAI could not perform, yet the model still attempted to produce solutions in a considerable number of cases. Similar patterns appeared in tasks requiring interpretation of device states or configuration settings.

These findings indicate that the CAI does not consistently recognise when a task falls outside its capabilities.

The broader pattern suggests that the CAI distinguishes effectively between automatable and non-automatable cases only when the escalation cues are explicit in the text or when similar examples are well represented in the training data. When escalation depends on subtle contextual signals, variable workflows or real-time system information, the model's decisions become less predictable. This aligns with observations in prior research, which has shown that LLMs struggle to identify task boundaries when they cannot access the information required to validate their reasoning.

Overall, the findings indicate that the CAI system can provide reliable escalation behaviour in clearly structured situations, but its performance is not yet consistent enough to allow autonomous decision-making across all types of pharmacy IT support tasks. These limitations reinforce the need for a hybrid model in which CAI handles routine, well-defined queries while human support agents remain responsible for tasks that involve uncertainty, sensitive information or operational risk.

Research question 3. What types of first line in pharmacy IT support requests are suitable for automation using CAI?

The findings give a clear indication of which types of support tasks are suitable for automation in pharmacy IT customer support. The categories that performed the best in this study shared three characteristics. They are low in operational risk, they follow a stable and predictable workflow, and they depend on written instructions that remain consistent over time. This aligns with findings from previous LLM research, which shows that these models work best when they can rely on written information (Szymanski et al., 2024). In this study, tasks such as billing, report generation, device troubleshooting and routine navigation aligned closely with these conditions. These tasks do not depend on complex decision-making and often involve repeated support interactions, which makes them well suited for automation.

The results also showed that certain individual queries from categories that were not strong overall could still be automated. Password reset requests are a good example. These performed well even though the broader access and certificates category did not. Many tasks in that category still require human intervention under Pharmadata's current procedures, but the password reset workflow is stable and does not rely on backend checks. This suggests that even in more challenging categories, narrow and well-defined routines can be automated when the underlying process is consistent.

The testing phase provided a detailed picture of how the Copilot Studio CAI agent behaves in practice. The agent produced clear and well-structured answers in these categories, and it used correct terminology even when the original ticket wording varied. This behaviour corresponds with the strengths identified in related LLM studies, where models tend to perform well in summarisation, instruction following, and question answering that rely on explicit textual cues (Patil & Gudivada, 2024). In this sense, the CAI’s most successful tasks resembled the “everyday AI” use cases described in prior studies, meaning routine interactions based on stable written procedures. They include explaining forms, guiding users through multi step menus, and clarifying how to access features inside Pharmadata’s platforms Omapd and pd3. In the results, these tasks were often completed in one or two prompts. The short prompt chains indicate that the model recognised user intent rapidly and did not require extensive refinement before producing a suitable response. Some low-risk categories also revealed useful side benefits. For example, the CAI was able to rephrase unclear problem descriptions into clearer steps, which was an essential action needed to solve the case. This mirrors the role LLMs have played in other sectors where information retrieval are central tasks. While this study did not include classification or clustering as explicit use cases, several responses demonstrated that the CAI could reorganise fragmented ticket information into coherent summaries. These behaviours suggest opportunities for future support tools that combine CAI with pre classification functions.

Despite the strong performance in routine tasks, several categories exposed clear limits. Issues involving access rights, certificate validation, or interactions across multiple systems could not be automated safely. These cases required backend knowledge that the CAI did not have based on the historical ticket data. The model attempted to infer missing details based on user text, but these inferences were often incorrect or incomplete. This pattern follows earlier research where LLMs struggle with factual consistency when they lack grounding in up-to-date system information. In pharmacy IT, this gap becomes critical, because many tasks involve verifying the state of external devices or confirming whether a change has been applied. The CAI also struggled with medication sensitive topics where accuracy is essential. The model occasionally provided accurate sounding instructions that did not meet safety expectations, which confirms that these tasks must remain under human control. In addition to these technical limits, the success of each category depended on the quality and stability of the underlying workflow. Functional categories that have frequent changes depending on the pharmacy were also more difficult for the CAI to handle. The model relies heavily on consistent patterns in the training material. When such patterns were weak, the model’s responses became less reliable. This reinforces the importance of

mapping use cases before deployment. Automation should be introduced only in categories where the procedure is stable and documented.

Considering these results, it is evident that CAI can address a meaningful subset of first line pharmacy IT support, but only in clearly defined areas. The tasks most ready for automation are those where written instructions already drive human work. These tasks offer efficiency gains without compromising safety. Tasks that require system checks, clinical judgement, or access modifications remain unsuitable for CAI and should be supported through hybrid models where the AI provides drafting or information retrieval but not final guidance.

Research question 4. What system-level structures and requirements are necessary to integrate a CAI-based support agent into existing pharmacy IT workflows?

The findings of this thesis indicate that successful integration of a CAI system into pharmacy IT support depends on a combination of technical foundations, organisational readiness, and regulatory alignment. These factors are interdependent, and the evaluation suggests that weaknesses in any one of them would limit the reliability or safe use of the CAI system.

From a technical standpoint, the results emphasise the need for a robust and up-to-date knowledge base. In the evaluation, the CAI operated only on historical ticket data, which restricted its ability to provide precise guidance in queries that required current procedures or rare problem scenarios. These limitations were visible in accuracy and helpfulness scores, particularly in categories where workflows change over time or vary between pharmacies. This indicates that integration will depend on the organisation's ability to provide structured, reliable and broad dataset sources that the system can draw from. Without such infrastructure the model would continue to rely on short-sighted knowledge, resulting in inconsistent guidance in practice.

The organisational context is equally significant. The evaluation was conducted in a controlled environment and did not include actual customer users. The results suggest that integration will require a phase of internal testing where support agents assess the system's behaviour, refine boundaries for safe automation, and identify task types where the CAI adds value. This process is not only technical but also cultural: the introduction of CAI changes how cases are triaged, how information flows, and how responsibility is shared. Training should cover both technical use and role boundaries, so that staff understand when to trust the AI's suggestion and when to override it.

Regulatory requirements form the third area for consideration. Pharmacy IT support involves personal data, authentication processes, and workflows connected to medication handling. These domains sit within the scope of

GDPR and the EU AI Act, which classify such systems as high risk (European Commission, 2024). The results underline the relevance of these constraints, as the CAI's weakest performance appeared in safety-sensitive categories. This implies that integration must include safeguards such as transparent logging, traceable decision pathways, and clear boundaries that prevent the CAI from taking actions affecting access rights, medication-related steps, or configurations without human oversight. Compliance is therefore not an external formality, but an operational necessity shaped directly by the system's performance characteristics.

Considered as a whole the integration of CAI into pharmacy IT support is possible and promising, but not a plug and play exercise. It requires alignment between technical architecture, business goals, and regulatory frameworks. The results suggest that CAI could become a stable component of first-line support when these conditions are met, improving efficiency and consistency while ensuring that human experts retain control over safety-critical decisions. The working hypothesis of this thesis was that CAI use would be most suitable for low-risk, single-system cases in which users mainly seek confirmation or clear procedural guidance. The findings largely support this hypothesis. The CAI system performed most consistently in routine, well-documented workflows that matched these conditions, while tasks involving system-state checks, multi-system dependencies or medication-related logic remained unsuitable for automation and required human oversight. At the same time, the results highlight that even within low-risk categories, reliable use of CAI still depends on strengthened escalation rules and a more robust data foundation. Overall, the hypothesis can therefore be considered confirmed for the clearly defined subset of first-line tasks identified in this study, but not generalisable to all pharmacy IT support cases.

5.2 Recommendations

These recommendations should be interpreted in the context of Pharmadata's ongoing strategic efforts to improve the efficiency, availability, and accessibility of its customer support operations. Rather than proposing a standalone technological change, the findings of this study support a gradual and proactive development path in which CAI is used to strengthen existing support practices. The first step in introducing CAI into Pharmacy IT customer support is to define what the system is meant to achieve. As this study has shown, CAI can assist with many routine support tasks, but every answer must still be read with appropriate caution. The model's uneven accuracy makes it clear that CAI should support existing practices rather than replace them. There are promising situations where the tool can improve efficiency, but these benefits appear only when expectations are realistic and when shared guidelines direct its use. Stakeholders need a common understanding of the situations where CAI can support work and the situations where human oversight remains essential. Decisions about data management,

documentation standards, and technical integration must be made in advance before deployment.

5.2.1 Establishing safe and effective CAI-supported workflows

The evaluation results indicate that CAI aligns well with Pharmadata's strategic direction in ongoing digital transformation, which emphasises improving support efficiency through digital tools. The CAI model's strongest performance appeared in areas where terminology, workflows, and instructions are stable. These include billing, reporting, device troubleshooting, and routine navigation in pd3 and Omapd, the core pharmacy IT platforms used across Pharmadata's customer base. These findings suggest that CAI can contribute meaningfully to service efficiency by handling questions that currently consume a large share of first-line support time. Early adoption in these areas would allow service desk personnel to focus on more complex cases that require diagnostic reasoning or system-state checks. A gradual rollout would allow Pharmadata to build competence around CAI-supported workflows. With clear boundaries and transparent oversight, the technology can strengthen internal knowledge-sharing and help standardise support quality across agents. Over time, CAI could become an integral component of Pharmadata's long-term service strategy. However, to reach this point the system must operate within well-defined constraints and be supported by reliable data sources and consistent escalation procedures.

Robust escalation boundaries are a prerequisite for CAI-supported workflows to operate safely, predictably, and consistently across different task types. Although the CAI correctly escalated many tasks related to certificates and complex configurations, its behaviour was less reliable in medication-related categories, particularly dose dispensing. These inconsistencies highlight the need for strengthened escalation boundaries. Three recommendations follow from the results of this thesis:

- 1. Strengthen medication-related escalation rules**
Any uncertainty in tasks involving dose dispensing, prescription handling, or other medication-sensitive workflows should result in immediate escalation. The model demonstrated uneven judgement in these contexts, and even more strict boundaries are therefore necessary.
- 2. Expand the clarity rule**
The evaluation showed that unclear queries often led to assumptions or incomplete answers rather than a request for clarification. The CAI should escalate whenever user intent remains uncertain after one clarifying attempt. This protects against situations where confident but inaccurate instructions would mislead the user.
- 3. Escalate all tasks requiring system-state verification**
Queries that depend on logs, device conditions, certificate states,

network connectivity, or permission checks cannot be handled safely without access to real-time information. As long as the CAI system does not have access to live system states, these tasks should be redirected to human support.

Embedding these rules into the CAI workflow ensures that the system remains fast and helpful in routine tasks while avoiding unsafe autonomy in sensitive areas. This structure supports a hybrid division of labour in which CAI handles stable interactions and human agents manage tasks requiring contextual interpretation or operational checks.

In addition to operational suitability, CAI-supported workflows should also be evaluated from an efficiency and resource perspective. While this study focuses on accuracy, escalation behaviour, and task suitability, the practical value of CAI lies in its potential to reduce routine first-line support workload. Future implementation phases should assess CAI's impact on support efficiency, for example through changes in drafting time, average handling time, or resolution rates in reliable task categories. These metrics can be translated into cost estimates to assess potential savings or capacity gains and to prioritise CAI use cases with clear operational and economic value.

5.2.2 Strengthening data and technical implementation

The Findings in Chapter 4 highlight that many of the CAI's accuracy issues arise not from weaknesses in the model itself but from the limitations of the underlying training data. Historical support tickets allowed the CAI to learn recurring phrasing and common workflow patterns, but they did not represent the full range of rare, evolving, or safety-sensitive tasks. Moreover, the ticket dataset did not always present a single correct or standardised way of resolving an issue. This was because in several categories, similar problems had been handled through multiple different actions. This variability meant that the model could not identify a stable procedural pattern at times, which in turn reduced its ability to produce consistent and context-appropriate guidance. Strengthening the system's data foundation is therefore essential for improving reliability, safety, and predictability. Two recommendations follow from this:

To begin with, a broader and more heterogeneous dataset should be developed and used. Prior research on domain adaptation consistently shows that expanding both the size of the model and the diversity of its training data yields stronger performance on specialised tasks (Patil & Gudivada, 2024). Studies on scaling have shown that when language models are trained with more parameters and more pre-training data at the same time, their performance improves in a predictable way across different benchmark tasks (Patil & Gudivada, 2024). In this context, the historical ticket set should be expanded both in terms of time and volume, so that it covers a longer period of

support activity and encounters a wider range of examples. A more extensive dataset would also increase the likelihood of capturing rare, complex, or low-frequency cases, reducing the risk that the CAI learns only the most common patterns. Expanding the available data in these ways would expose the CAI to a broader spectrum of real support scenarios and provide more stable procedural patterns for the model to learn from.

In addition, the CAI model should be connected to maintained and standardised documentation sources such as Pharmadata's documentation source Omaneuvo. Historical tickets offer valuable insight into real support interactions, but they do not contain the authoritative procedures, rules, and exceptions required for dependable guidance. Linking the CAI to up-to-date system information would improve accuracy by grounding the model in verified procedures instead of relying solely on past conversations. If technically feasible, giving the CAI controlled access to relevant live system information, such as device states or certificate validity, would further reduce uncertainty in categories where correct actions depend on current system conditions. Together, these measures would create a more robust data foundation that supports both accuracy and consistent behaviour across pharmacies.

The evaluation also points to several opportunities for long-term scalability. Although the CAI is not yet ready for fully autonomous operation, it already performs consistently in tasks driven by stable terminology and well-documented workflows. These strengths create a foundation for gradual expansion of the CAI's role within Pharmadata's service ecosystem. A staged adoption strategy would allow the organisation to build operational competence and refine governance structures before exposing the system to customer-facing use. Several practical design choices can guide this approach.

A practical starting point is a controlled deployment model in which all customer-facing messages generated by the system require human verification before being sent. In this configuration, the CAI drafts responses, retrieves relevant information, and summarises procedural steps, but a support agent remains responsible for reviewing and approving the content. This safeguard is particularly important in categories where incorrect guidance could affect authentication processes, system access, or medication-related workflows. This approach offers balance between efficiency and safety. It allows the company to benefit immediately from reduced drafting time and improved consistency in routine cases, while ensuring that customers are not exposed to inaccurate instructions during the early stages of adoption. It also provides a structured environment in which support agents can evaluate the CAI's behaviour across different categories, identify recurring points of failure and refine escalation logic as the system matures. Over time in categories where

performance becomes consistently reliable, the level of mandatory human review could be reassessed.

Furthermore, the CAI should be integrated directly into the existing service desk workflow rather than introducing a separate tool. Embedding the system within the interface already used by support agents, would facilitate seamless handover during escalation. When the CAI is built into the same environment, agents can review drafts, approve messages, or take over an interaction without switching applications. This reduces friction and supports efficient case handling. Automatic tracking of all CAI-generated actions is an essential element of this integration. Comprehensive logs enable support agents to track how a response was produced, provide material for internal training, and allow supervisors to analyse recurring issues. Logging also supports auditability, which is required for compliance with GDPR and the EU AI Act (European Commission, 2024). Overall, these features help create a workflow where AI assistance enhances everyday operations while maintaining continuity, transparency, and accountability across all support activities.

Looking ahead, collaboration with an external technical partner should be considered. Developing and maintaining an enterprise CAI system requires specialised expertise in model governance, secure system integration, and scalable infrastructure, which can be difficult to sustain internally. A partnership model would allow Pharmadata to focus on domain-specific rules, safe workflow design, and human oversight, while technical implementation and maintenance are handled by specialists. This division of responsibilities supports long-term reliability and ensures that the CAI evolves in a controlled and transparent manner. It would also be sensible to compare several providers and underlying CAI models before committing to a long-term solution. Vendors differ in their integration capabilities, support arrangements, security features, and pricing structures. A systematic comparison would clarify which provider can meet Pharmadata's technical and regulatory needs and what the long-term financial implications would be. This evaluation is essential for determining whether the project can scale sustainably and remain operationally viable over time.

By combining a strengthened data foundation with secure technical integration and systematic human oversight, CAI can evolve into a scalable and dependable component of Pharmadata's customer support environment. This staged approach provides a clear and controlled pathway for expanding the system's role while preserving the levels of safety, accuracy, and compliance required in pharmacy IT.

5.3 Critical reflections on CAI in pharmacy IT support

5.3.1 Limitations of CAI

The literature used in this study reflects a field that is developing at a rapid pace. Most of the publications were recent and many were produced in response to the growing interest in CAI. This pace of publishing shows strong engagement across the research community, yet it also raises questions about the durability and reliability of current findings. Several papers noted their own limits, such as restricted resources for model testing or the lack of transparency in the tools they examined. Much of the existing work focuses on models like ChatGPT, even though the internal configurations of these systems are not publicly available. This lack of transparency makes it difficult to compare insights from the broader literature with the behaviour of enterprise tools, which might operate within different technical, security, and regulatory conditions.

The empirical material in this thesis brings its own limitations that affect both the reliability of the findings and generalisability of the results. The analysis relied on historical support tickets from Pharmadata, which means that the dataset mirrors one organisation's processes, documentation habits, and internal terminology. These tickets document actual pharmacy IT problems, but they do not represent the full range of issues that support teams face each day. Many real cases develop across several messages where the customer clarifies symptoms, adds missing steps, or changes the description of the problem as the conversation progresses. Support staff often resolve these cases through iterative questions, small corrections, and trial-and-error steps. Their suggestions may not work on the first attempt, and they adjust the guidance as new information becomes available. The quality of the historical support tickets also affects the strength of the results. Some categories contained clear and consistent steps, while others included missing details or several possible solutions. When the documentation lacked clarity, the model performed less reliably. This shows that the evaluation measured both the capability of the model and the clarity of the material it was asked to interpret. In practice the model can only perform as well as the information it receives. The prompts used in this study were based on real problems that customers had previously reported. They were created from the historical tickets, and the wording was adjusted to remove identifiable information and not be identical from the historical ticket training data. Although the phrasing changed, the technical issues remained the same as in the original cases, and each prompt reflected the content of an actual support request. However, the prompts did not fully capture the conversational nature of real support interactions. Although they were based on authentic cases, the testing environment did not reflect the ongoing clarifications, changes in problem descriptions, or the evolving dialogue that often occurs in live

situations. As a result, the evaluation represents the underlying problems accurately but does not replicate the full complexity of real-time support. This limits how confidently the findings can be applied to actual customer interactions.

Human assessment introduces uncertainty. Even with defined scoring criteria, the evaluation of clarity, accuracy, and escalation logic relies on human judgement. The scoring focused on how well the model identified the correct solution based on the information available in the historical tickets. This creates a methodological limitation because the ratings reflect an interpretation of the most appropriate resolution rather than an absolute measure of correctness. In several support categories, multiple solution paths may be valid, particularly in workflows where troubleshooting involves alternative sequences of steps or case-specific adjustments. These nuances are difficult to capture through a single scoring approach. The evaluation might favour answers that resemble the organisation's standard practices and unintentionally penalise answers that are technically correct but expressed differently. Historical tickets represent the organisation's preferred way of solving problems, and this may not cover all acceptable or technically feasible resolutions. A solution can be correct, but if it does not follow the familiar internal method, it may still be rated as wrong or incomplete. This introduces a potential bias toward solutions that mirror the organisation's documented behaviour, which should be acknowledged when interpreting the results. A broader panel of evaluators would have increased reliability of the results. Multiple raters would allow for comparison across different interpretations of correctness and escalation appropriateness, especially in cases where the underlying issue can be approached in more than one way. This would also make it possible to compare how consistently different evaluators rate the same answers, which helps improve the reliability of qualitative studies. This was not possible within the scope of the present study, which limits the generalisability of the scoring outcomes.

The testing in this study was conducted in Copilot Studio, which offers a controlled and secure environment for evaluating CAI behaviour without exposing sensitive healthcare-related information. The model had no access to live system states, time-dependent processes, or diagnostic logs. These elements are often central to understanding and resolving pharmacy IT issues, since many problems emerge only when specific configurations change or when certain events occur in a particular order. Because of these limitations, the results describe how the model behaves with static prompts rather than how it would perform when interacting with dynamic system behaviour. This distinction is important, since real support work depends heavily on context that evolves over time and cannot be fully represented through fixed test inputs. It is also possible that the model's performance would have been stronger if

it had access to real system information, logs, and configuration data. The absence of these elements should therefore be recognised as a limiting factor when interpreting the results.

The fast pace of development in LLMs adds another layer of complexity to the interpretation of the results. Depending on the model used, safety mechanisms, response behaviour, and language support can shift through updates that are not always visible to end users. Such changes may affect performance in different application domains and are particularly important for smaller languages such as Finnish, where improvements tend to appear gradually. During the writing of this thesis, a new generation of LLMs was released, which further illustrates how rapidly the underlying technology evolves. A different or more recent model might therefore produce different performance outcomes than those observed in this study. This means that some weaknesses observed in the evaluation may become less prominent over time, while new behaviours may also emerge as model providers refine their systems. The findings should therefore be understood as a snapshot of the model's performance during the period of testing. Any practical use of CAI in pharmacy IT will require continuous monitoring, repeated testing, and adjustment to ensure that system behaviour remains stable and reliable as the underlying technology keeps advancing.

5.3.2 Ethical considerations

Pharmacy IT involves sensitive information and strict regulatory requirements, which place clear limits on the use of CAI. This study used anonymised material and conducted all testing in a controlled setting. Any real deployment would require strong protections to ensure that personal or medication-related data does not reach non-approved systems. Users would also need guidance on what types of information is safe to be included in prompts. This is essential, since CAI tools do not always distinguish between harmless details and information that could reveal patient identity, clinical history, or other confidential aspects of pharmacy operations. Even within an enterprise environment, automated tools must be configured so that they do not expose sensitive data or replicate content from one user's interaction in responses to another user. This must be enforced through strict technical controls and governance procedures. Without these safeguards, the use of CAI in pharmacy IT could create unacceptable privacy risks. Robust data handling policies, clear user guidance, and controlled access are therefore critical components of any ethically acceptable deployment.

Interpretability is another central concern when deploying CAI in pharmacy IT. The model can express answers in a confident and fluent way even when the reasoning behind those answers is difficult to understand or verify. This

lack of transparency becomes ethically significant because incorrect or partially incorrect guidance can influence medication workflows, user access rights, or reimbursement-related processes. These areas require high levels of reliability and the consequences of mistakes can extend beyond simple technical inconvenience. For this reason, it would be important that human supporters remain responsible for final decisions until the CAI model can be fully trusted with these actions. Automated suggestions can be useful as decision support, but they cannot replace professional judgement in situations where patient safety or system integrity might be affected. Categories that involve dose dispensing, certificate handling, or other clinically sensitive functions require well-defined escalation rules to ensure that such cases are always reviewed by human experts. This safeguards against the risk of over-reliance on automated systems in contexts where precision and accountability are essential.

Finally, there are also organisational factors that shape the ethical use of CAI in pharmacy IT. The introduction of AI tools can create uncertainty among support staff. This is particularly evident regarding changes in roles, decision-making authority, and expectations for technical competence. These concerns are relevant because the effectiveness of CAI depends not only on technical performance but also on how well the organisation manages the transition. In this context, the primary purpose of CAI is to assist customer support workers rather than replace them. Its value lies in reducing repetitive workload and improving consistency, while responsibility for critical decisions remains with human staff. To achieve this, organisations need clear communication strategies that explain the intended role of the technology and its limitations. Staff must also have opportunities to interact with the tool in a controlled and low-risk environment. These conditions help build trust, support learning, and encourage realistic expectations about the capabilities of CAI in daily support work.

5.3.3 Directions for future research

The limitations identified in this study open several possibilities for future research on the use of CAI in pharmacy IT customer support:

1. Comparative evaluation of different CAI models

This study examined only one CAI model, which limits conclusions about the broader suitability of CAI tools in pharmacy IT support. Future research should compare the performance of several models with different architectures, safety controls, and domain adaptation strategies. General-purpose models, enterprise-specific systems, and domain-fine-tuned models may each offer distinct advantages. A comparative approach would help identify which model types perform best in high-risk categories such as dose

dispensing, certificate management, and medication-related workflows. It would also reveal which limitations are specific to a single model and which are inherent to CAI technology more broadly. These comparisons would give organisations a clearer evidence base when choosing an AI system for first-line support and help technology providers understand where additional safety mechanisms or domain adaptations are needed.

2. Long-term evaluation of CAI performance in evolving pharmacy IT environments

Pharmacy IT systems do not remain static but change over time. Medication rules evolve, reimbursement processes change, and technical systems undergo regular updates. At the same time, CAI models themselves change through shifts in safety layers, language support, and backend infrastructure. Because of this there is a need for long-lasting research that tracks CAI performance over time. Such studies could examine whether the accuracy of CAI responses remains stable when system updates change workflows or when regulatory changes introduce new requirements for medication safety and user authentication. They could also monitor how CAI behaviour shifts after model provider updates, even if the underlying version number does not change. Long-term monitoring would help organisations define appropriate re-evaluation intervals, identify emerging risks, and maintain safe deployment over time. This is particularly relevant in pharmacy IT, where incorrect guidance can have system-wide or clinical implications.

3. Multi-input CAI for handling screenshots and interface-based troubleshooting

A significant part of pharmacy IT support concerns information that cannot be conveyed well through text alone. Error codes, device displays, barcode scanner screens or interface states are often shared as screenshots. Future research should examine the feasibility of multi-input CAI systems that can analyse images alongside text. Key research questions include whether the model can reliably recognise error messages, identify configuration states, or distinguish between similar interface elements. This would bring CAI closer to real-world troubleshooting, where visual information plays a central role. Multi-input capabilities may also improve the model's ability to escalate complex cases, since the system would have a clearer understanding of when the situation involves medication safety or technical risks. Research in this area would extend the practical utility of CAI tools and reduce the burden on human support staff, who currently rely heavily on visual diagnostics.

These directions for future work show the scientific value of this study by offering a structured basis for evaluating CAI in regulated digital environments and by identifying clear boundaries for safe automation. The findings also carry clinical relevance, as they clarify which workflows can be supported without compromising medication safety. The results further indicate potential economic benefits by reducing repetitive workload and improving service

efficiency when CAI is used in stable, high-volume tasks. On a broader societal level, the study demonstrates how well-governed CAI can strengthen trust in digital health services and support more reliable access to pharmacy IT systems.

Looking ahead, the integration of CAI into pharmacy IT support points toward a gradual shift toward hybrid human and AI workflows. As models improve and become more tightly aligned with domain requirements, new opportunities for safe automation and more adaptive support solutions will emerge. This progression opens promising avenues for research and suggests that CAI can become a responsible and impactful element in the future of digital pharmacy services.

References

- Adnan, M., Asghar, Z., Rizwan, M., Khan, T.F., Umer, N. & Hussain, I. (2025) A hybrid AI chatbot framework for intelligent pharmacy management systems. *Spectrum of Engineering Sciences*, 3(5), pp. 919–928.
- Aldoseri, A., Al-Khalifa, K.N. & Hamouda, A.M. (2024) AI-powered innovation in digital transformation: Key pillars and industry impact. *Sustainability*, 16(5), p. 1790.
- Almeman, A. (2024) The digital transformation in pharmacy: Embracing online platforms and the cosmeceutical paradigm shift. *Journal of Health, Population and Nutrition*, 43(1). <https://doi.org/10.1186/s41043-024-00550-2>
- Alnefaie, A., Singh, S., Kocaballi, B. & Prasad, M. (2021) An overview of conversational agents: Applications, challenges and future directions. *Proceedings of the 17th International Conference on Web Information Systems and Technologies*. Scitepress.
- Andrade, I. & Tumelero, C. (2022) Increasing customer service efficiency through artificial intelligence chatbot. *Revista de Gestão*, 29(3), pp. 238–251. <https://doi.org/10.1108/REG-07-2021-0120>
- Badar, M.A., Gupta, R., Srivastava, P., Ali, I. & Cudney, E.A. (eds.) (2024) *Handbook of digital innovation, transformation, and sustainable development in a post-pandemic era*. CRC Press.
- Chang, Y. et al. (2024) A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), pp. 1–45.
- Coman, M.M. & Kifor, C.V. (2024) The emerging and disruptive technologies – a risk-based approach. *Land Forces Academy Review*, 29(2), pp. 237–246.
- Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A. & Sheikh, A. (2011) The case study approach. *BMC Medical Research Methodology*, 11, p. 100. <https://doi.org/10.1186/1471-2288-11-100>
- Dehalwar, K.S.S.N. & Sharma, S.N. (2024) Exploring the distinctions between quantitative and qualitative research methods. *Think India Journal*, 27(1), pp. 7–15.
- Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G. & Beck, H.P. (2003) The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6), pp. 697–718.

- Egala, S.B., Nyarku, K.M., Quansah, F. & Boateng, F. (2024) Digital transformation in an emerging economy: Exploring organizational drivers. *Cogent Social Sciences*, 10(1). <https://doi.org/10.1080/23311886.2024.2302217>
- European Commission (2024) Artificial intelligence in healthcare. Available at: https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en (Accessed 4 November 2025).
- Feng, C.M., Botha, E. & Pitt, L. (2024) From HAL to GenAI: Optimizing chatbot impacts with CARE. *Business Horizons*, 67(5), pp. 537–548.
- Habbal, F., Kolmos, A., Hadgraft, R.G., Holgaard, J.E. & Reda, K. (2024) Reshaping engineering education: Addressing complex human challenges. Springer Nature.
- Hassija, V., Chakrabarti, A., Singh, A., Chamola, V. & Sikdar, B. (2023) Unleashing the potential of conversational AI: Amplifying Chat-GPT's capabilities and tackling technical hurdles. *IEEE Access*, pp. 1–1. <https://doi.org/10.1109/ACCESS.2023.3339553>
- Huang, Y.K., Hsieh, C.H., Li, W., Chang, C. & Fan, W.S. (2019) Preliminary study of factors affecting the spread and resistance of consumers' use of AI customer service. *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*, pp. 132–138.
- Huang, L. et al. (2025) A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), pp. 1–55.
- Krakowski, S., Luger, J. & Raisch, S. (2023) Artificial intelligence and the changing sources of competitive advantage. *Strategic Management Journal*, 44(6), pp. 1425–1453.
- Kraus, S., Jones, P., Kailer, N., Weinmann, A., Chaparro-Banegas, N. & Roig-Tierno, N. (2021) Digital transformation: An overview of the current state of the art of research. *SAGE Open*, 11(3), p. 21582440211047576.
- Laymouna, M., Ma, Y., Lessard, D., Schuster, T., Engler, K. & Lebouché, B. (2024) Roles, users, benefits, and limitations of chatbots in health care: Rapid review. *Journal of Medical Internet Research*, 26, e56930. <https://doi.org/10.2196/56930>
- Li, J. et al. (2024) User experience design professionals' perceptions of generative artificial intelligence. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–18. <https://doi.org/10.1145/3613904.3642114>

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. & Neubig, G. (2023) Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), pp. 1–35.

McTear, M. & Ashurkina, M. (2024) A new era in conversational AI. In: *Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents*. Berkeley, CA: Apress, pp. 1–16.

Naveed, H. et al. (2023) A comprehensive overview of large language models. arXiv. <https://doi.org/10.48550/arxiv.2307.06435>

Nguyen, K.D. (2023) A behavioural decision-making framework for agent-based models. PhD thesis. Utrecht University. <https://doi.org/10.33540/472>

Norman, D.A. (2013) *The design of everyday things: Revised and expanded edition*. New York: Basic Books.

Ogundipe, A., Sim, T.F. & Emmerton, L. (2024) Prescription for digital evolution: Transformative recommendations for pharmacy practice in the digital age. *Journal of Pharmacy Practice*, 38(2), pp. 237–248. <https://doi.org/10.1177/08971900241277049>

Omol, E.J. (2024) Organizational digital transformation: From evolution to future trends. *Digital Transformation and Society*, 3(3), pp. 240–256.

Pasas-Farmer, S. & Jain, R. (2025) From discovery to delivery: Governance of AI in the pharmaceutical industry. *Green Analytical Chemistry*, 13, p. 100268.

Patel, R. & Jain, A. (2021) Empathy and nuance in customer support chatbots. *Journal of AI Research*, 45(11), pp. 981–999.

Patil, R. & Gudivada, V. (2024) A review of current trends, techniques, and challenges in large language models (LLMs). *Applied Sciences*, 14(5), p. 2074.

Peltoniemi, T., Suomi, R., Peura, S. & Lähteenoja, M.N. (2021) Electronic prescription as a driver for digitalization in Finnish pharmacies. *BMC Health Services Research*, 21(1), p. 1017.

Pharmadata Oy (2025) Company webpage. Available at: <https://pharmadata.fi/yritys/> (Accessed 20 October 2025).

Qu, Y., Huang, S., Li, L., Nie, P. & Yao, Y. (2025) Beyond intentions: A critical survey of misalignment in LLMs. *Computers, Materials & Continua*, 85(1).

Ris, M. & Puvača, N. (2024) Digitisation and digitalisation in pharmaceutical workflows. *Journal of Pharmacy Informatics*, 6(2), pp. 89–102.

Sankar, B. & Sen, D. (2025) A novel idea generation tool using a structured conversational AI (CAI) system. *AI EDAM*, 39, p. e11.

Shin, E., Hartman, M. & Ramanathan, M. (2024) Performance of the ChatGPT large language model for decision support in community pharmacy. *British Journal of Clinical Pharmacology*, 90(12), pp. 3320–3333.

Shirahama, N., Nakaya, N. & Watanabe, S. (2024) Comparative study of visual analogue scale and Likert scale using chat generation AI. *Proceedings of the 11th International Conference on Intelligent Systems and Image Processing*, pp. 195–202.

Singh, S.U. & Namin, A.S. (2025) A survey on chatbots and large language models: Testing and evaluation techniques. *Natural Language Processing Journal*, p. 100128.

Sterbini, A. & Temperini, M. (2024) Open-source or proprietary language models? An initial comparison on the assessment of an educational task. *Proceedings of the 21st International Conference on Information Technology Based Higher Education and Training (ITHET)*, pp. 1–7. <https://doi.org/10.1109/ITHET61869.2024.10837616>

Subramonyam, H., Pea, R., Pondoc, C., Agrawala, M. & Seifert, C. (2024) Bridging the gulf of envisioning: Cognitive challenges in prompt-based interactions with LLMs. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Szymanski, A. et al. (2024) Integrating expertise in LLMs: Crafting a customized nutrition assistant with refined template instructions. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–22. <https://doi.org/10.1145/3613904.3641924>

Troshin, S., Mohammed, W., Meng, Y., Monz, C., Fokkens, A. & Niculae, V. (2025) Control the temperature: Selective sampling for diverse and high-quality LLM outputs. *arXiv preprint arXiv:2510.01218*.

Varzaru, A.A. & Bocean, C.G. (2024) Digital transformation and innovation: The influence of digital technologies on turnover from innovation activities and types of innovation. *Systems*, 12(9), p. 359. <https://doi.org/10.3390/systems12090359>

Verhoef, P.C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J.Q., Fabian, N. & Haenlein, M. (2021) Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*, 122, pp. 889–901.

Wahl, J., Müller, R. & Zimmermann, T. (2024) The evolution of digital pharmacy in Germany: Opportunities and regulatory challenges. *European Journal of Health Policy*, 18(2), pp. 112–127.

Wong, A., Flanagan, T., Covington, E.W., Nguyen, E., Linn, D., Brummel, G., Hoffmaster, B., Isaacs, D. & Kane-Gill, S.L. (2025) Forecasting the impact of artificial intelligence on clinical pharmacy practice. *Journal of the American College of Clinical Pharmacy*, 8(4), pp. 302–310.

Wu, T., Terry, M. & Cai, C.J. (2022) AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–22.

Zafar, A., Parthasarathy, V.B., Chan, L.V., Shahid, S., Khan, A.I. & Shahid, A. (2024) Building trust in conversational AI: A review and solution architecture using large language models and knowledge graphs. *Big Data and Cognitive Computing*, 8(6), p. 70. <https://doi.org/10.3390/bdcc8060070>

Zeb, S., Nizamullah, F.N.U., Abbasi, N. & Fahad, M. (2024) AI in healthcare: Revolutionizing diagnosis and therapy. *International Journal of Multidisciplinary Sciences and Arts*, 3(3), pp. 118–128.

Appendix A. Evaluation Framework for Copilot Performance

Date & Time:
Query number:
Query description:
Summary of Copilot response:
Evaluation notes <ul style="list-style-type: none"> • Strengths: • Weaknesses: • Escalation Decision (Correct / Incorrect / Not applicable)

Evaluation criteria

Score scale: (1: Strongly disagree, 2: Disagree, 3: Neither agree nor disagree, 4: Agree, 5: Strongly agree)

Criteria	Description	Score
Clarity	The answer is easy to read and understand. It avoids confusing or unclear wording.	
Accuracy	The information provided is correct and relevant to the question.	
Helpfulness	The answer provides a useful solution or helps move the issue forward.	
Structure	The answer is well organised and presented in a logical way.	
Terminology	The answer uses the correct professional and technical terms for the pharmacy IT context.	
Escalation logic	The model recognises when it cannot solve the issue and correctly suggests escalation to a human support agent.	
Amount of Prompts		

Appendix B. Copilot CAI performance evaluation data

Functional category	Query number	Clarity	Accuracy	Helpfulness	Structure	Terminology	Escalation logic	Amount of prompts	Average of all evaluation criteria
Billing	1	5	4	5	5	4	5	1	4.67
	2	5	5	5	5	4	5	2	4.83
	3	5	3	2	4	4	1	1	3.17
	4	4	3	2	4	4	2	2	3.17
	5	4	5	4	5	5	5	2	4.67
	6	4	4	5	4	5	5	2	4.50
	7	4	3	3	4	4	4	2	3.67
	8	3	3	4	4	5	5	3	4.00
	9	5	5	4	5	5	4	2	4.67
	10	4	4	4	5	4	5	1	4.33
	11	4	3	2	4	4	5	1	3.67
	12	4	4	3	4	4	4	3	3.83
	13	5	4	2	4	3	5	2	3.83
	14	4	4	4	4	5	5	1	4.33
	15	4	3	4	4	4	4	5	3.83
	16	4	3	3	4	4	3	4	3.50
	17	5	5	4	5	5	5	1	4.83
	18	4	5	4	4	5	4	2	4.33
	19	4	4	2	4	4	4	1	3.83
Average in billing		4.26	3.89	3.47	4.32	4.32	4.26	2.00	4.09
Prescription Handling	20	4	5	5	4	4	5	2	4.50
	21	4	4	2	4	4	5	1	3.83
	22	5	4	3	4	5	3	2	4.00
	23	5	2	2	4	4	3	4	3.33
	24	5	1	2	5	5	2	1	3.33
	25	5	2	2	5	5	3	2	3.67
	26	5	5	4	5	4	5	2	4.67
	27	4	4	4	4	5	3	2	4.00
	28	4	5	3	5	5	5	2	4.50
	29	4	5	4	4	5	4	1	4.33
	30	4	5	5	4	5	5	1	4.67
Average in prescription handling		4.45	3.82	3.27	4.36	4.64	3.91	1.82	4.08
Inventory Management	31	5	5	4	4	4	5	2	4.50
	32	4	5	3	4	5	5	1	4.33
	33	4	3	3	4	4	2	3	3.33
	34	4	3	3	4	5	3	3	3.67
	35	5	3	3	4	4	3	2	3.67
	36	4	4	4	4	4	4	3	4.00
	37	4	3	2	4	5	3	1	3.50
	38	5	4	4	5	4	4	2	4.33
	39	4	3	2	4	4	3	2	3.33
	40	4	3	3	4	4	4	1	3.67
	Average in inventory management		4.30	3.60	3.10	4.10	4.30	3.60	2.00
Device Issues	41	3	4	4	5	4	5	4	4.17
	42	5	4	4	5	5	4	2	4.50
	43	4	2	2	4	4	2	1	3.00
	44	4	4	4	4	4	4	3	4.00
	45	5	3	3	5	4	3	2	3.83
	46	4	3	4	5	4	5	3	4.17
	47	4	3	4	4	4	5	2	4.00
	48	5	5	5	5	4	5	1	4.83
	49	5	4	4	4	4	5	2	4.33
	50	5	5	4	4	4	5	2	4.50
	51	4	2	2	4	3	2	1	2.83
Average in device issues		4.36	3.55	3.64	4.45	4.00	4.09	2.09	4.02
Miscellaneous	52	4	4	3	3	4	5	1	3.83
	53	4	5	5	4	5	5	1	4.67
	54	4	5	4	4	4	5	1	4.33
	55	4	1	1	3	3	2	2	2.33
	56	4	3	2	4	4	5	1	3.67
	57	4	3	3	4	5	4	2	3.83
	58	4	5	5	4	5	5	1	4.67
	59	5	4	4	5	5	5	2	4.67
	60	4	4	3	4	5	3	1	3.83
	61	5	5	3	4	5	5	1	4.50
	62	5	5	5	5	4	5	2	4.83
	63	5	5	3	4	4	5	1	4.33
	64	4	4	4	5	5	4	1	4.33
	65	4	5	3	4	4	5	1	4.17
	66	3	3	2	4	4	2	2	3.00
	67	4	2	2	3	4	2	1	2.83
	68	4	3	2	3	4	4	1	3.33
	69	4	3	3	4	4	3	2	3.50

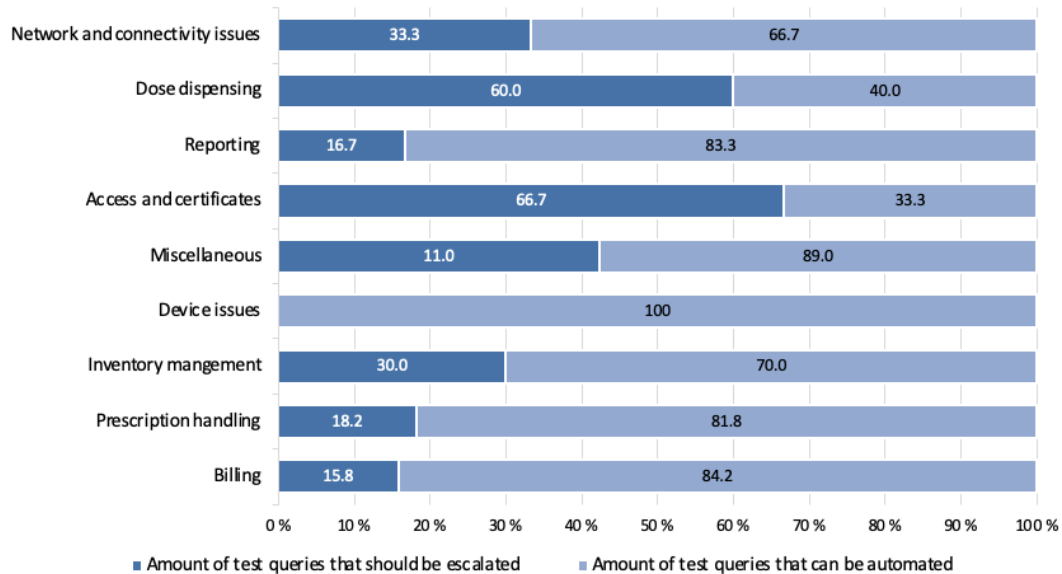
	70	4	2	2	3	4	2	2	2.83
	71	4	3	2	3	4	4	1	3.33
	72	4	4	2	4	4	4	1	3.67
	73	4	4	2	4	5	5	1	4.00
	74	4	4	4	5	5	4	3	4.33
	75	4	5	3	3	5	5	1	4.17
	76	3	4	2	4	4	4	1	3.50
	77	3	5	4	4	4	4	1	4.00
Average in miscellaneous		4.04	3.85	3.00	3.88	4.35	4.08	1.35	3.87
Access and Certificates	78	4	2	2	4	3	2	1	2.83
	79	4	4	3	3	4	5	2	3.83
	80	3	4	5	4	3	5	2	4.00
	81	4	3	3	4	4	5	1	3.83
	82	4	3	2	4	4	5	1	3.67
	83	4	4	2	4	4	5	1	3.83
	84	4	3	2	4	4	5	1	3.67
	85	5	5	2	4	4	5	1	4.17
	86	4	5	4	5	4	5	2	4.50
Average in access and certificates		4.00	3.67	2.78	4.00	3.78	4.67	1.33	3.81
Reporting	87	4	3	4	5	4	5	3	4.17
	88	5	5	5	4	4	4	2	4.50
	89	4	5	4	4	4	5	2	4.33
	90	4	5	3	4	5	5	2	4.33
	91	5	4	4	5	4	5	3	4.50
	92	4	4	4	5	4	4	3	4.17
Average in reporting		4.33	4.33	4.00	4.50	4.17	4.67	2.50	4.33
Dose Dispensing	93	4	3	2	4	3	5	1	3.50
	94	5	4	4	4	4	4	2	4.17
	95	4	5	4	4	4	2	2	3.83
	96	4	4	4	4	4	2	4	3.67
	97	4	2	2	4	3	2	1	2.83
Average in dose dispensing		4.20	3.60	3.20	4.00	3.60	3.00	2.00	3.60
Network and Connectivity Issues	98	4	5	4	4	5	5	2	4.50
	99	4	4	2	4	4	5	2	3.83
	100	3	1	3	4	4	3	1	3.00
Average in category network and connectivity issues		3.67	3.33	3.00	4.00	4.33	4.33	1.67	3.78
Average in all functional categories		4.20	3.78	3.25	4.16	4.23	4.09	1.78	3.95

Appendix C. Distribution of test queries requiring escalation versus automation

Functional category	Amount of test queries that should be escalated	Amount of test queries that can be automated
Billing	3	16
Prescription handling	2	9
Inventory mangement	3	7
Device issues	0	11
Miscellaneous	11	15
Access and certificates	6	3
Reporting	1	5
Dose dispensing	3	2
Network and connectivity issues	1	2

Functional category	% of test queries that should be escalated	% of test queries that should be automated
Billing	15.79 %	84.21 %
Prescription handling	18.18 %	81.82 %
Inventory mangement	30.00 %	70.00 %
Device issues	0.00 %	100.00 %
Miscellaneous	11.00 %	89.00 %
Access and certificates	66.67 %	33.33 %
Reporting	16.67 %	83.33 %
Dose dispensing	60.00 %	40.00 %
Network and connectivity issues	33.33 %	66.67 %

Distribution of test queries requiring escalation vs automation



Appendix D. Evaluation of escalation decisions by category and query

Query number	Category	Should the query be escalated to a human support agent	Was it escalated correctly
1	Billing	no	yes
2	Billing	no	yes
3	Billing	no	no
4	Billing	no	no
5	Billing	no	yes
6	Billing	no	yes
7	Billing	no	yes
8	Billing	no	yes
9	Billing	no	yes
10	Billing	no	yes
11	Billing	yes	yes
12	Billing	no	yes
13	Billing	Yes	yes
14	Billing	no	yes
15	Billing	no	yes
16	Billing	no	yes
17	Billing	no	yes
18	Billing	no	yes
19	Billing	yes	yes
20	Prescription Handling	no	yes
21	Prescription Handling	yes	yes
22	Prescription Handling	no	no
23	Prescription Handling	no	yes
24	Prescription Handling	no	no
25	Prescription Handling	no	no
26	Prescription Handling	no	yes
27	Prescription Handling	no	yes
28	Prescription Handling	yes	yes
29	Prescription Handling	no	yes
30	Prescription Handling	no	yes
31	Inventory Management	no	yes
32	Inventory Management	yes	yes
33	Inventory Management	yes	yes

34	Inventory Management	yes	yes
35	Inventory Management	no	yes
36	Inventory Management	no	yes
37	Inventory Management	no	no
38	Inventory Management	no	yes
39	Inventory Management	no	no
40	Inventory Management	no	yes
41	Device Issues	no	yes
42	Device Issues	no	yes
43	Device Issues	no	no
44	Device Issues	no	yes
45	Device Issues	no	yes
46	Device Issues	no	yes
47	Device Issues	no	yes
48	Device Issues	no	yes
49	Device Issues	no	yes
50	Device Issues	no	yes
51	Device Issues	no	no
52	Miscellaneous	yes	yes
53	Miscellaneous	no	yes
54	Miscellaneous	no	yes
55	Miscellaneous	no	no
56	Miscellaneous	yes	yes
57	Miscellaneous	no	yes
58	Miscellaneous	no	yes
59	Miscellaneous	no	yes
60	Miscellaneous	no	no
61	Miscellaneous	no	yes
62	Miscellaneous	no	yes
63	Miscellaneous	yes	yes
64	Miscellaneous	no	yes
65	Miscellaneous	yes	yes
66	Miscellaneous	no	no
67	Miscellaneous	no	yes

67	Miscellaneous	no	yes
68	Miscellaneous	yes	yes
69	Miscellaneous	yes	no
70	Miscellaneous	no	no
71	Miscellaneous	yes	yes
72	Miscellaneous	yes	yes
73	Miscellaneous	yes	yes
74	Miscellaneous	no	yes
75	Miscellaneous	yes	yes
76	Miscellaneous	yes	yes
77	Miscellaneous	no	yes
78	Access and Certificates	no	no
79	Access and Certificates	Yes	yes
80	Access and Certificates	no	yes
81	Access and Certificates	Yes	yes
82	Access and Certificates	Yes	yes
83	Access and Certificates	yes	yes
84	Access and Certificates	Yes	yes
85	Access and Certificates	yes	yes
86	Access and Certificates	no	yes
87	Reporting	no	yes
88	Reporting	no	yes
89	Reporting	no	yes
90	Reporting	yes	yes
91	Reporting	no	yes
92	Reporting	no	yes
93	Dose Dispensing	yes	yes
94	Dose Dispensing	no	yes
95	Dose Dispensing	Yes	no
96	Dose Dispensing	Yes	no
97	Dose Dispensing	no	no
98	Network and Connectivity Issues	no	yes
99	Network and Connectivity Issues	yes	yes
100	Network and Connectivity Issues	no	yes