Resume

Digitale løsninger til seksuel sundhed rummer et betydeligt potentiale for at øge unges adgang til information og rådgivning i situationer, hvor emnet opleves som følsomt, stigmatiseret eller vanskeligt at adressere gennem traditionelle sundhedstilbud. Samtidig er området karakteriseret ved høje krav til anonymitet, datasikkerhed og medicinsk korrekthed, og tidligere forskning peger på, at manglende tillid ofte udgør en central barriere for anvendelsen af digitale sundhedssystemer. Især inden for seksuel sundhed kan uklare og upræcise svar have direkte betydning for brugerens tryghed og beslutninger. Dette speciale undersøger derfor, hvordan og i hvilket omfang en seksual sundheds-chatbot opfattes som troværdig af unge, når den er designet med fokus på sikkerhed, anonymitet, personlig interaktion og medicinsk korrekthed i en særligt sensitiv sundhedskontekst, samt hvordan denne oplevede troværdighed forholder sig til andre etablerede digitale systemer inden for seksuel sundhed.

Arbejdet bygger videre på et forudgående eksplorative pre-thesis-projekt, som havde til formål at afdække muligheder og udfordringer ved digitale interventioner inden for seksuel sundhed. Pre-thesis-arbejdet omfattede et systematisk litteraturstudie samt kvalitative interviews med domæneeksperter fra danske organisationer inden for seksuel sundhed. Gennem dette arbejde blev der identificeret centrale problemstillinger relateret til kapacitetsbegrænsninger i eksisterende rådgivningstilbud, unges behov for anonym, diskret og lettilgængelig adgang til information, samt fagpersoners bekymringer omkring ansvar, kvalitet og begrænsning i digitale løsninger. Disse indsigter bidrog til en afklaring af, hvor teknologiske løsninger kan supplere, men ikke erstatte, menneskelig rådgivning, og dannede grundlag for formuleringen af konkrete designkrav og afgrænsninger i nærværende speciale.

Tidligere studier inden for Human-Computer Interaction (HCI), digitale sundhedssystemer og conversational agents beskriver tillid som et komplekst og flerdimensionelt fænomen, hvor brugeres vurderinger ikke kan reduceres til én samlet tillidsopfattelse eller til systemets tekniske kapabiliteter alene. I stedet forstås tillid som sammensat af flere distinkte, men relaterede dimensioner, som tilsammen former, hvordan sundhedsteknologier vurderes som troværdige. Den gennemgåede litteratur peger samtidig på, at sådanne flerdimensionelle tillidsforhold ikke kan indfanges gennem simple mål for tilfredshed eller brugsintention, men kræver evalueringsmetoder, der eksplicit måler og sammenligner flere tillidsdimensioner. Dette speciale positionerer sig i forlængelse af denne tilgang ved at anvende en flerdimensionel forståelse af tillid som analytisk ramme for en komparativ evaluering af forskellige digitale systemer i en realistisk brugskontekst.

På baggrund af resultater fra pre-thesis og tidligere forskning blev der designet og implementeret en prototype, *SafeBubble*, som en sikker og medicinsk forankret seksual sundheds-chatbot. Systemet er udviklet som en on-premises løsning baseret på retrieval-augmented generation (RAG), hvor svar genereres med udgangspunkt i udvalgte, medicinsk validerede danske kilder. Denne arkitektur muliggør gennemsigtighed og ansvarlighed i svarene og understøtter samtidig anonymitet og lokal databehandling. Systemet er bevidst designet til både at kunne give forklarende og støttende svar og til eksplicit at markere usikkerhed eller behov for professionel rådgivning i situationer, hvor digital vejledning ikke er tilstrækkelig. Brugergrænsefladen er holdt enkel og genkendelig for at minimere visuel påvirkning af tillidsvurderinger og fokusere evalueringen på systemets sikkerhed, adfærd og svar.

Evalueringen blev gennemført som et between-subject design med tre systembetingelser: *SafeBubble*, et etableret dansk seksual-sundhedswebsite (*Sex & Samfund*) og en generel conversational AI (*ChatGPT*). Deltagerne blev tilfældigt fordelt på systemerne og gennemførte en række opgavebaserede informationssøgninger, der afspejlede realistiske og hverdagsnære behov inden for seksuel sundhed. Efter interaktionen besvarede deltagerne et spørgeskema, hvor tillid blev operationaliseret som et flerdimensionelt begreb baseret på etablerede tillidsrammer (Ability, Benevolence og Integrity) samt supplerende trust-in-automation-konstrukter relateret til blandt andet forståelighed, pålidelighed og adfærdsintention. Data blev analyseret ved hjælp af komparative statistiske metoder med henblik på at undersøge forskelle mellem systembetingelserne på tværs af tillidsdimensioner.

Resultaterne viser, at deltagernes vurderinger af tillid varierer på tværs af både systemer og dimensioner, hvilket indikerer, at tillid ikke kan betragtes som et entydigt eller samlet fænomen. *SafeBubble* vurderes særligt højt på dimensioner relateret til integritet og velvilje, samt på oplevet pålidelighed og forståelighed, hvilket peger på, at et system, der eksplicit er designet med fokus på medicinsk forankring, ansvarlig afgrænsning og anonymitet, kan opfattes som troværdigt i en sensitiv sundhedskontekst. Samtidig observeres der ikke signifikante forskelle på alle tillidsdimensioner, herunder oplevet evne og generel tillid til automation, hvilket understreger, at højere vurderinger på enkelte dimensioner ikke nødvendigvis generaliseres til en samlet tillidsvurdering.

Endvidere viser resultaterne, at forskelle i oplevet troværdighed ikke automatisk omsættes til øget adfærdsintention. Dette indikerer en delvis afkobling mellem vurderingen af et systems troværdighed og deltagernes umiddelbare intention om at handle på eller fortsætte brugen af systemet. I diskussionen fortolkes dette som et udtryk for, at adfærdsintention i en seksual-sundhedskontekst kan være påvirket af yderligere faktorer såsom personlig relevans og kontekst og ikke alene af oplevet tillid.

Samlet set peger evalueringen på, at en seksual sundheds-chatbot kan opfattes som troværdig på udvalgte dimensioner, uden at dette nødvendigvis fører til ændret adfærd eller øget brug. I den komparative evaluering viser resultaterne endvidere, at *SafeBubble* på flere centrale tillidsdimensioner, herunder velvilje, integritet, forståelighed og oplevet pålidelighed, vurderes på niveau med eller højere end både et etableret domænespecifikt informationssystem og en generel conversational AI. Specialets resultater bidrager dermed med empirisk viden om, hvordan tillid manifesterer sig forskelligt på tværs af dimensioner og systemer, og understøtter en forståelse af tillid som et kontekstuelt og flerdimensionelt begreb. Afslutningsvis peger specialet på flere retninger for fremtidigt arbejde, herunder tættere og mere systematisk inddragelse af medicinsk ekspertise, forbedring af sproglig kvalitet og retrieval-kontrol, samt udvidede tillidsmålinger relateret til anonymitet og datasikkerhed. Endvidere fremhæves behovet for langvarige evalueringer, der kan belyse, hvordan troværdighed udvikler sig over tid og ved gentagen brug.

Would You Trust a Chatbot With Your Sex Life? Implementing and Evaluating a Secure, Anonyms Chatbot for Young People

Frederikke Filtenborg Michaelsen
fmicha20@student.aau.dk
Aalborg University
Denmark

ABSTRACT

Trust is a central precondition for the adoption of digital interventions in sexual health, where users are asked to rely on technology in contexts characterized by sensitivity, stigma, and personal risk. While conversational agents are increasingly proposed as scalable solutions for sexual-health guidance, limited empirical work has examined how such systems are perceived in terms of trustworthiness, particularly in comparison to established alternatives. This study investigates to what extent a sexual-health chatbot can be perceived as trustworthy by young people, and how it compares to other established systems. Building on prior research and expert-informed requirements, a secure, medically grounded chatbot was designed using an on-premises, retrieval-augmented architecture that prioritizes anonymity, data security, transparency, and responsible scoping. The chatbot was evaluated in a between-subjects study against two comparison systems: an established sexual-health information website and a general-purpose conversational AI. Trust was assessed using a multidimensional questionnaire grounded in established trust and trust-in-automation frameworks, and analyzed through comparative statistical methods. The findings show that the chatbot can be perceived as trustworthy across several trust-related dimensions, particularly those related to benevolence, integrity, reliability, and understandability. However, trust did not manifest uniformly across all dimensions, nor did higher trust scores translate into differences in behavioral intention. Rather than establishing causal effects of specific design choices, the study demonstrates that a sexual-health chatbot implemented within a privacy-preserving and medically grounded system context can achieve trust levels comparable to, and in some respects exceeding, those of existing alternatives. These results contribute empirical insight into trust formation in conversational sexual-health interventions and highlight the importance of architectural transparency, interaction quality, and contextual evaluation when deploying AI systems in sensitive health domains.

1 INTRODUCTION

Sexually transmitted infections (STIs) among young people in Denmark have reached epidemic levels. Why this situation has escalated to such severity, and why infection rates continue to rise, are questions that health professionals still struggle to answer [24, 39, 44]. These medical questions and research are beyond the scope of this study and are left to medical experts. Instead, this study examines how digital technologies can support existing initiatives aimed at controlling the spread of STIs among young people in Denmark.

Previous research has shown that several meaningful initiatives already exist within the public health sector - both through the Danish healthcare system and through non-governmental organizations leading preventive and educational efforts. Two organizations in particular, *Sex & Samfund* and *Checkpoint (AIDS-Fondet)*, are recognized as national front runners in STI prevention and receive governmental support for their work [2, 40]. Their initiatives include free testing at clinics, counseling, educational websites, and the distribution of free STI home-testing kits. However, these organizations face practical constraints, as their capacity and outreach depend heavily on state funding. This limited capacity opens an opportunity to explore whether technology could strengthen existing efforts through new digital interventions.

Building on prior exploratory work conducted in a pre-thesis project, *Digital Interventions for STIs: Identifying Areas and Opportunities for Design* [24], this project extends an initial investigation that mapped key areas where technology could support STI prevention among young people in Denmark. The pre-thesis identified several opportunities for digital interventions, particularly within counseling, partner notification, and home-testing support, and established an empirical foundation for further research. Building on these findings, this project investigates the potential of a chatbot as a digital intervention to support existing STI prevention practices. A chatbot solution is particularly suitable in this context, as it can meet key requirements related to personalized and interactive solutions. The chatbot can complement existing initiatives provided by organizations such as *Sex & Samfund*, for example by supporting their email-based counseling services, and to extend in-person counseling offered by *Checkpoint* in situations where capacity is limited. Through conversational interaction, the chatbot can provide immediate, low-threshold guidance that aligns with how young people already seek information in digital contexts. Given the sensitive nature of sexual-health information, the solution must meet strict requirements regarding security, anonymity, and medical accuracy. At the same time, the use of a conversational system raises an important question: will young people trust a digital system with such sensitive and private information?

These considerations lead to the central research question of this project:

> *To what extent can a sexual-health chatbot be perceived as trustworthy by young people, and how does it compare to other established systems?*

In this study, established systems refer to existing digital sources for sexual-health information and support, specifically the institutional platform *Sex & Samfund* and the widely used general-purpose conversational system *ChatGPT*.

This project argues for why a chatbot designed to support existing preventive practices constitutes a valid and relevant intervention. While the study examines the design and evaluation of

such a solution, particular attention is given to the technical implementation of a secure, anonymous, and medically accurate chatbot. This aspect is essential for assessing whether the intervention is technically feasible and whether the chosen system architecture can realistically support the required levels of privacy, security, and medical credibility. The solution will be evaluated both by domain experts, who represent the professional context, and by young people, the primary target group, to determine whether the research question can be answered.

2 RELATED WORK

This section reviews prior work relevant to the present study across four interrelated areas. First, research on digital sexual-health interventions is examined to establish the broader public-health context and to identify recurring challenges related to credibility, engagement, and privacy. Second, existing work on sexual-health chatbots is reviewed with a particular focus on reported limitations, adoption barriers, and trust-related concerns in sensitive health contexts. Third, research on trust in digital interventions and AI-based health communication is discussed to establish the theoretical foundations for treating trust as a multidimensional construct and for evaluating trust empirically in chatbot systems. Finally, findings from the pre-thesis are summarized to highlight empirically identified gaps in current practice and to motivate the focus of the present study.

2.1 Digital Interventions for Sexual-Health and STIs

Sexually transmitted infections (STIs) remain a persistent public health challenge, particularly among young people, where incidence rates continue to be high despite long-standing prevention efforts. In Denmark, surveillance data show that chlamydia is the most frequently reported bacterial STI, with the majority of diagnosed cases occurring among individuals aged 15-29 years [39, 44]. Epidemiological studies further indicate substantial under diagnosis, as many infections are asymptomatic, contributing to ongoing transmission and delayed treatment [10]. These patterns underline the need for interventions that can reach young populations early, lower barriers to testing, and support preventive behaviors.

In response to these challenges, a growing body of research has explored digital interventions as means of improving access to sexual-health information, testing, and care. Prior work has investigated web-based platforms and mobile applications aimed at increasing STI knowledge, promoting safer sexual behavior, and supporting partner notification and testing practices [29, 50]. Many of these interventions take the form of informational websites, symptom checkers, or mobile applications that provide general guidance on STI prevention and testing pathways. While such systems generally report positive effects on knowledge acquisition and self-reported behavioral intentions, they are often characterized by largely static content and limited interaction. As a result, they tend to function as informational supplements rather than interactive support tools. These studies also highlight that uptake and sustained use vary considerably across user groups, pointing to broader socio-technical factors influencing engagement.

Several studies specifically address the role of mobile health applications in sexual-health contexts. Empirical investigations of STI-related apps show that while awareness and prior use are relatively limited, there is pronounced interest among younger users, particularly when apps are perceived as medically reliable and endorsed by trusted institutions [14]. Existing applications commonly focus on providing basic information about STIs, guidance on testing locations, reminders for regular testing, and in some cases partner notification support. Features such as anonymity and ease of access are consistently ranked as valuable, whereas concerns related to data security and credibility remain important barriers to adoption [14, 15]. Similar findings are reported in other empirical studies, which suggest that these applications are typically perceived as useful supplements to clinical consultations rather than replacements for professional care [5, 47].

At the same time, prior research consistently raises concerns regarding the credibility and medical accuracy of existing digital sexual-health interventions. Reviews and empirical studies point out that a substantial portion of publicly available sexual-health information, particularly in app-based and online formats, lacks adequate medical validation, contains outdated recommendations, or presents simplified guidance that may be misleading in complex situations [9, 14, 18, 29]. Many existing solutions rely on manually curated content that is infrequently updated or drawn from multiple, diverse sources without transparent validation. This lack of medical reliability is highlighted as a critical limitation, as users may struggle to distinguish trustworthy sources from unregulated content, especially when seeking information under time pressure or emotional distress [9]. Consequently, several studies argue that future digital interventions must place stronger emphasis on medically vetted content, explicit sourcing, and institutional accountability in order to be considered legitimate components of sexual-health care.

Beyond issues of accuracy, prior literature also emphasizes that effective digital sexual-health interventions must move beyond static information delivery. Multiple studies argue that personalization and interactivity are central to supporting meaningful engagement and relevance for users, particularly when addressing situational concerns such as potential exposure, symptoms, or partner communication [29, 50]. However, many existing systems offer only limited adaptation, typically through predefined pathways or FAQ-style navigation. Interventions that fail to adapt to individual contexts or allow users to explore questions interactively are often perceived as insufficiently responsive, limiting their usefulness in moments of uncertainty. This has led researchers to call for systems that can support more dialog and context-sensitive forms of interaction, while still operating within medically responsible boundaries [18, 28, 29, 50].

Research has also examined how digital sexual-health interventions intersect with broader changes in sexual practices and media use. The widespread adoption of dating and geosocial networking applications has been associated with increased sexual risk behaviors, particularly among men who have sex with men, but has simultaneously motivated the development of digital tools for risk communication and partner notification [14, 50]. Existing technological responses include notification systems embedded in dating platforms, external partner-notification tools, and information campaigns delivered through the same channels where sexual encounters are initiated. This dual role of technology, as both a contributor

to risk and a potential vehicle for prevention, has shaped much of the recent research agenda, motivating interventions that seek to meet users within the digital environments they already inhabit.

Finally, studies consistently highlight anonymity and data security as foundational requirements for digital interventions in sexual-health. Given the stigma and sensitivity associated with STIs, users' willingness to engage with digital systems is closely tied to perceptions of privacy, confidentiality, and protection of personal data [14, 47]. Many existing systems attempt to address these concerns by allowing anonymous access or by minimizing data collection. However, prior work shows that perceived anonymity is fragile and easily undermined by unclear data practices or opaque system architectures. Interventions that fail to convincingly address these concerns risk exclusion by the very user groups they aim to support. As such, prior work suggests that future digital sexual-health systems must be designed not only to provide accurate and personalized information, but also to operate within environments that users perceive as secure and anonymous.

Despite promising results, prior work consistently emphasizes that technological availability alone does not ensure meaningful value. Reviews of digital sexual-health interventions note that long-term impact is often difficult to establish, and that many systems struggle to achieve sustained engagement beyond initial use [18, 29]. Taken together, this body of work suggests that while digital interventions hold clear potential for addressing persistent sexual-health challenges, their success depends on credibility, interactivity, and perceived safety-factors that become particularly critical when users are asked to rely on digital systems in sensitive health contexts.

2.2 Challenges and Limitations of Sexual-health Chatbots

Although chatbots are frequently positioned as scalable channels for sexual-health information and service navigation, prior work shows that their adoption is not a "given" - especially in contexts marked by stigma, embarrassment, and highly sensitive disclosures. Studies focusing specifically on sexual-health chatbots show that perceived advantages such as anonymity and ease of access coexist with clear concerns about competence, interpersonal limitations, and safety. For example, participants in a sexual-health chatbot study described how the ability to ask questions without being identified could lower the barrier for discussing embarrassing topics, particularly compared to face-to-face consultations or general practitioners without sexual-health specialization [32]. At the same time, the same study reports barriers tied to perceived chatbot limitations: reduced confidence in chatbot competence for meaningful consultations, constrained conversational depth, and concerns that responses can feel generic, which is especially problematic in a domain requiring sensitive language and contextual nuance [32].

Previous research on healthcare chatbots shows that users are often hesitant to rely on AI-based systems, particularly because of concerns about accuracy, data security, and the lack of human understanding. In a mixed-methods study, participants described worries about receiving incorrect advice, uncertainty about how secure their data would be, and the chatbot's limited ability to respond in an empathetic way. Survey results from the same study showed only moderate overall acceptability, with higher acceptance among users who perceived the chatbot as trustworthy and useful [30]. Importantly, the qualitative findings suggest that this hesitation was not abstract or theoretical. Participants explicitly linked their concerns to not knowing where the chatbot's information came from, doubts about how mature or reliable the technology was, and fears that incorrect or incomplete advice could cause harm. In sexual-health contexts, these concerns appear even more pronounced, as users are often not just seeking general information but trying to manage personal risk. As a result, incorrect guidance may be perceived as having direct and personal consequences rather than being a minor inconvenience [30, 32].

Previous review studies of healthcare chatbots show that many proposed systems appear promising, but that the empirical evidence supporting their safe and reliable use is limited. A systematic review of healthcare conversational agents with unconstrained user input found that most published evaluations were narrow in scope and that patient safety was rarely assessed, even though these systems were intended to provide health-related guidance [18]. Similarly, a broader review of conversational agents in health reports rapid growth in the field, but also points to persistent gaps related to user engagement, interaction quality, and how well systems support different user needs [28]. Taken together, these reviews suggest that while healthcare chatbots are often presented as promising solutions, existing research has frequently not examined the factors that are most critical for whether users are willing to rely on such systems in practice. In particular, prior work has paid limited attention to users' ability to assess the medical correctness of chatbot responses, to understand the system's limitations, and to judge whether it can handle personal and context-dependent questions. In sensitive domains such as sexual-health, additional factors further shape reliance, including perceived anonymity, confidence in data security, and the overall quality of the interaction when users disclose intimate concerns. When these aspects are insufficiently addressed, users may hesitate to trust the system or may avoid using it altogether, even if the chatbot appears useful at surface level [18, 28].

Evidence from mental-health chatbots illustrates both the potential of conversational agents and how quickly trust can break down when interaction quality is poor. In a randomized controlled trial of the Woebot mental-health chatbot, high user engagement and short-term reductions in depressive symptoms were reported. Qualitative feedback indicated that features such as regular check-ins and a perceived empathetic tone contributed positively to users' experiences [8]. At the same time, users described negative experiences when the chatbot failed to understand unexpected inputs, repeated the same responses, or entered conversational loops. These breakdowns were interpreted as signs that the system lacked competence and reliability, leading to frustration and reduced trust [8].

For a sexual-health chatbot, similar interaction failures are likely to have more serious consequences. When users ask personal or sensitive questions, misunderstandings or generic responses may be interpreted not merely as usability issues, but as indications that the system cannot safely handle complex or intimate situations. As a result, such breakdowns may discourage users from disclosing relevant details or from continuing to use the system at all [8, 32].

Taken together, prior studies suggest that the central challenge in deploying sexual-health chatbots is not simply whether they can provide information, but whether users consider them trustworthy enough to use in practice. Trust determines whether users are willing to share sensitive information, follow guidance, and rely on the system when making decisions under uncertainty. Across the literature, several recurring barriers to trust and adoption are identified, including uncertainty about medical accuracy, lack of transparency regarding information sources, concerns about confidentiality and data security, limited ability to respond empathetically, and shallow or repetitive conversations [8, 31, 32]. In sexual-health contexts, these barriers are particularly critical, as anonymity, protection of personal data, and a sense of safety strongly influence whether users engage with a system in the first place.

2.3 Trust in Digital Interventions and AI Health Communication

Research in human-computer interaction and automation has consistently examined trust as a central factor influencing whether users are willing to rely on technical systems under conditions of uncertainty and potential risk. Prior work conceptualizes trust in automation as a user's willingness to rely on a system despite uncertainty about its behavior or outcomes. From this work, empirical observations across automated domains demonstrate that inappropriate trust calibration can undermine system effectiveness in two ways: users may over-rely on systems that should be treated with caution (misuse), or avoid systems that could otherwise provide meaningful support (disuse) [20]. This body of work establishes trust as a practical condition for use, rather than a purely abstract attitude.

Further research on automated decision-support systems shows that trust plays a decisive role in whether users choose to rely on system outputs, disclose information, or follow recommendations. Such decisions are typically made under uncertainty, as users often lack insight into how systems function, what data they rely on, or how reliable their outputs are [11]. As a result, trust becomes a key determinant of whether a chatbot is used as intended or rejected altogether.

The same studies further demonstrates that trust in technology differs from interpersonal trust. Unlike interpersonal trust, which often involves assumptions about intentions or moral character, trust in technical systems is primarily shaped by perceptions of competence, reliability, and functional purpose [11, 20]. This distinction is supported by information systems research, where studies show that users evaluate technologies based on whether systems behave consistently, function as expected, and appear aligned with user interests rather than hidden agendas [23]. These findings are particularly relevant for chatbots, which may be experienced as conversational partners while remaining technical systems. This distinction matters because conversational interaction can create expectations of understanding or authority that exceed actual system capabilities, increasing the risk of misplaced trust [11, 20].

Empirical studies of AI-based systems further show that trust is dynamic and context-dependent. Prior work reports that users may initially trust systems based on early interactions or surface-level cues, but that trust can deteriorate rapidly following perceived errors or breakdowns. Conversely, technically reliable systems may be underutilized when users do not understand how decisions are generated or feel uncomfortable relying on opaque processes [11, 20]. These findings highlight why trust cannot be inferred from technical performance alone and must be examined empirically.

In the context of health chatbots, prior studies link trust closely to perceptions of medical accuracy, transparency, and safety. Previous research shows that users are more willing to engage with health chatbots when information is perceived as medically grounded and when system limitations are clearly communicated [30]. Conversely, uncertainty about information sources, fear of receiving incorrect advice, or concerns about a system's ability to handle personal or nuanced questions undermine trust and acceptance [30]. Similarly, research on conversational agents demonstrates that interaction breakdowns, such as repetitive responses or failure to respond appropriately, are interpreted as indicators of low competence and reliability, leading to reduced trust and disengagement [8].

Privacy, security, and anonymity further shape trust in digital health interventions. Empirical studies show that users' willingness to disclose personal information depends not only on perceived usefulness, but also on confidence that personal data will be protected. Prior work reports that concerns about data security, hacking, or inappropriate data use can significantly reduce trust and acceptance of health chatbots, even when users acknowledge their potential benefits [30]. In sensitive health domains, anonymity has been identified as a key enabler for engagement, as it reduces stigma and fear of judgment. However, prior research also shows that anonymity only supports trust when it is perceived as genuine and supported by credible data-handling practices.

Recent empirical work on trust in AI systems further shows that users' trust can be influenced by interaction cues that are not directly related to actual system reliability. Prior studies report that conversational features such as tone, fluency, and perceived empathy can increase users' comfort and willingness to engage with AI systems, even when system competence is limited [1]. Similarly, research on socially expressive conversational agents demonstrates that users may interpret emotionally expressive behavior as signals of trustworthiness [4]. Across these studies, such cues are found to facilitate initial engagement, but also raise concerns about over reliance when users place confidence in systems beyond their intended scope.

To address these challenges, prior research typically treats trust as a multidimensional concept rather than a single, overall attitude. Trust is thus understood not as one unified feeling, but as a combination of different beliefs that together shape whether users are willing to rely on a system.

Within organizational trust research, trust has been conceptualized as consisting of three core dimensions: ability (perceived competence), benevolence (perceived intention to act in the user's interest), and integrity (perceived adherence to acceptable principles and consistency of behavior) [21]. Building on this framework, information systems research demonstrates that these dimensions can be operationalised and measured reliably, and that they contribute differently to users' willingness to rely on a system [22]. This work shows that users may trust a system in some respects while remaining skeptical in others.

Within HCI and automation research, these trust dimensions have been adapted to account for the characteristics of technical systems. Rather than focusing on human intentions, studies emphasize factors such as reliability, predictability, understandability, and perceived system purpose as central components shaping trust in automation [17, 20]. Several studies further show that the relative importance of these dimensions varies depending on context, perceived risk, and task sensitivity, meaning that different aspects of trust become salient in different situations [11, 20].

This multidimensional view of trust has important implications for empirical evaluation. Instead of measuring trust as a single outcome, prior studies commonly use structured questionnaires that capture distinct but related components, such as trusting beliefs, trusting intentions, and general propensity to trust [17, 22]. This approach allows researchers to identify cases of partial or conditional trust, for example when users perceive a system as useful or competent but remain cautious about safety, transparency, or reliability. In sensitive health contexts, such distinctions are particularly important, as users' willingness to disclose personal information or follow system guidance may depend on specific trust dimensions rather than on overall trust alone.

Finally, research on trust measurement in AI systems suggests that trust and distrust should not be treated as opposite ends of a single scale. Prior work shows that users may simultaneously express moderate trust in a system's usefulness while maintaining strong reservations about its safety or reliability, particularly in complex or high-stakes domains [42]. This perspective supports examining trust empirically in specific contexts rather than assuming that positive engagement implies full trust.

Taken together, existing research establishes trust as a central but fragile factor in the adoption and use of digital and AI-based health interventions. While chatbots may lower barriers to engagement, trust can easily be undermined by uncertainty about accuracy, lack of transparency, privacy concerns, or poor interaction quality. These findings motivate further investigation into how users evaluate and place trust in concrete chatbot systems operating in sensitive health contexts, which directly informs the focus of the present study.

2.4 Findings From the Pre-thesis

The present study builds directly on findings from a pre-thesis [24] conducted by the author, which explored the role and limitations of existing digital initiatives within sexual-health through qualitative expert interviews. The purpose of the pre-thesis was to identify gaps in current practice and to examine where and how digital systems could meaningfully support sexual-health services, particularly for young people. Two domain experts were interviewed: one representing a counseling and testing service for sexually transmitted infections, and one representing a national organization working with sexual-health education and prevention [24].

Across both interviews, experts described a growing mismatch between demand and capacity in existing sexual-health services. Young people increasingly seek information and reassurance outside traditional clinical encounters, often motivated by urgency, uncertainty, or embarrassment. According to the experts, current counseling services are frequently unable to accommodate this demand in real time, resulting in long waiting times or missed opportunities for early guidance [24]. Digital interventions were therefore framed not as replacements for human counseling, but as potential entry points that could provide immediate, low-threshold access to medically grounded information.

At the same time, both experts highlighted significant shortcomings in current digital offerings. From the perspective of the counseling service, a key problem is that young users often arrive with information gathered from unregulated online sources, which can be misleading, incomplete, or anxiety-inducing [24]. This was described as an increasing burden on human counselors, who must first correct misconceptions before meaningful dialogue can take place. The expert emphasized that while anonymity is crucial for encouraging help-seeking behavior, it also increases the risk that users place undue trust in unreliable sources.

The expert from the sexual-health education organization similarly pointed to a lack of credible, coherent digital guidance that aligns with established medical practices and public health recommendations. While informational websites exist, they were described as poorly suited for users with situational or personalized questions, particularly in moments of concern related to potential exposure, symptoms, or partner communication [24]. According to the expert, current systems fail to support users in navigating uncertainty, often leaving them with either overly generic information or fragmented advice spread across multiple platforms.

Based on these expert perspectives, the pre-thesis identified 13 core requirements for future digital sexual-health systems. First, experts stressed the necessity of *medical legitimacy*, meaning that any digital intervention must clearly ground its responses in medically validated sources and communicate this grounding transparently to users [24]. Second, systems must support *anonymity* without sacrificing responsibility, ensuring that users feel safe to ask sensitive questions while also understanding the system's limitations [24]. Third, experts emphasized the importance of *appropriate scoping*, where digital systems provide guidance and orientation rather than diagnosis or treatment, and actively direct users to professional care when necessary [24].

Importantly, the pre-thesis also revealed that *trust* was not discussed by experts as a purely technical property, but as a relational and contextual concern [24]. Experts expressed concern that conversational digital systems may be perceived as authoritative sources of advice, leading users to rely on them in situations where human counseling or medical care is necessary [24]. This highlights a central tension: while digital systems may lower barriers to engagement, they simultaneously introduce new risks related to over reliance and misplaced trust.

Taken together, the pre-thesis establishes a clear motivation for the present study. While prior research and expert practice suggest strong potential for digital interventions in sexual-health, such as chatbots, they also underscore the need to understand whether users are willing to trust such systems, and under what conditions. This is particularly critical when interventions operate in domains characterized by sensitive data, stigma, and health-related uncertainty.

3 FROM PROBLEM TO CONCEPT

This project builds directly on the exploratory pre-thesis *Digital Interventions for STIs: Identifying Areas and Opportunities for Design*, which examined how digital technologies could support STI prevention and counseling through a combination of literature review and expert interviews with representatives from the organizations *Checkpoint* and *Sex & Samfund* [24]. Rather than proposing a single solution, the pre-thesis focused on identifying gaps in current practice and translating empirical insights into a set of concrete design requirements intended to guide future digital interventions. These requirements were summarized in a dedicated requirements table and discussed further in the pre-thesis section on future work and design directions [24].

Building on this foundation, the present study takes these requirements as a point of departure. Across expert interviews and literature analysis, several recurring requirements were identified. Focusing on these requirements allows the present study to move beyond a repetition of related work. While prior research outlines general challenges and opportunities in digital sexual-health, the requirements derived from the pre-thesis clarify what a viable intervention must support in practice. The present study therefore represents a transition from problem identification toward the selection and implementation of a concrete concept that explicitly addresses these requirements.

3.1 Explored Design Directions

Based on the requirements and future design directions outlined in the pre-thesis, several alternative concepts were considered during the early stages of this study [24]. Each design direction represented a different way of addressing the identified needs within sexual-health services.

One direction focused on redesigning anonymous partner-notification tools to improve usability and uptake, responding to expert concerns about the limited effectiveness of existing notification solutions. Another explored digital support for users of home STI testing, aiming to provide guidance before and after testing while reducing the manual counseling burden currently handled by *Sex & Samfund* [24]. Additional concepts included automated email-support systems and intelligent FAQ-style solutions intended to streamline information delivery, as well as broader personalized prevention platforms offering tailored behavioral guidance.

While each of these concepts addressed some of the requirements identified in the pre-thesis, they differed in how comprehensively they supported the full set of needs. Several concepts relied primarily on static information delivery, limiting their ability to support personalized, situation-specific questions. Others required extensive integration with existing health infrastructures or organizational workflows, shifting the focus away from user interaction and trust toward institutional implementation. This comparative exploration highlighted that meeting individual requirements in isolation was insufficient, instead, the concept needed to support broader coverage across multiple sexual-health areas, alongside anonymity, medical credibility, and personalized interaction [24].

3.2 Final Concept: A Secure Sexual-health Chatbot

The final concept selected for implementation is a secure, medically grounded chatbot designed to support sexual-health counseling in an anonymous and responsible manner. This choice is directly informed by the set of requirements identified in the pre-thesis, which were derived from expert interviews and a targeted literature review [24]. Rather than treating the chatbot as a generic solution, the concept is deliberately shaped by these requirements and by limitations of sexual-health chatbots reported in prior research.

Expert interviews conducted in the pre-thesis highlighted a persistent gap between demand and capacity in existing sexual-health counseling services [24]. This led to the requirement that a digital intervention should be able to extend access to medically grounded information without replacing human counselors. A chatbot satisfies this requirement by providing immediate, low-threshold access to guidance, while still allowing professional services to remain the primary point of care.

At the same time, experts emphasized that anonymity and data protection are essential for engagement in sexual-health contexts [24]. Prior research supports this observation, showing that anonymity can lower barriers to asking sensitive or embarrassing questions [32]. However, the same studies demonstrate that anonymity alone is insufficient. Users frequently express concerns about chatbot competence, limited conversational depth, and the risk of receiving generic or inappropriate responses, particularly in domains requiring sensitivity and contextual understanding [32]. These findings informed the requirement that any chatbot intervention must clearly communicate its scope and limitations and avoid presenting itself as an authoritative replacement for professional care.

Medical credibility and transparent sourcing were also identified as central requirements in the pre-thesis, driven by expert concerns about users relying on unregulated or misleading online information [24]. Empirical studies reinforce this concern, and show that uncertainty about information sources, fear of incorrect advice, and concerns about data security significantly reduce trust and acceptance of health chatbots [30]. Similarly, review studies report that many healthcare chatbots are deployed without sufficient evaluation of safety, interaction quality, or medical responsibility [18, 28]. In response, the present study treats medical grounding, transparent sourcing, and conservative scoping as one of the main requirements.

The pre-thesis further identified a strong need for personalized, situation-specific guidance [24]. Experts noted that existing websites and static information resources often fail to support users who are navigating uncertainty related to symptoms, exposure, or partner communication. While conversational interfaces are well suited to supporting personalized interaction, prior research cautions that poorly designed dialogue systems can quickly undermine trust. Furthermore, previous work demonstrated how interaction breakdowns, such as repetitive responses or conversational loops, are interpreted by users as signs of low competence and reliability, leading to disengagement [8]. In sexual-health settings, such breakdowns may have more serious consequences, as users are often managing perceived personal risk rather than general information needs.

Taken together, findings from the pre-thesis and prior research suggest that the central challenge is not whether a chatbot can be deployed, but whether a single digital intervention can satisfy the combined requirements of anonymity and data security, medical credibility, usability, and personalized interaction in a way that users perceive as trustworthy. On this basis, a chatbot was selected as the concept with the strongest potential to support these requirements simultaneously, as consistently emphasized across expert interviews and prior literature [24]. Rather than assuming that conversational interfaces are inherently appropriate, the concept was developed with explicit awareness of documented limitations of sexual-health chatbots. Insights from previous studies and expert concerns informed the design constraints applied to the prototype, ensuring that issues such as over reliance, shallow interaction, and unclear authority were actively addressed. The purpose of the prototype is therefore to empirically investigate whether a chatbot developed around these requirements can support trust among young users in a sensitive health context.

4 SYSTEM ARCHITECTURE

The architecture was designed to meet the requirements of a chatbot system described earlier in subsection 3.2, and builds upon the exploratory research conducted in the pre-thesis [24]. The system is intended for young people seeking information and advice about sexual-health, ranging from general questions to more personal concerns. The chatbot is not limited to simple fact-based questions but is also meant to supplement existing human counseling when the demand exceeds current capacity. This is achieved by supporting a dialog communication and follow up question from the chatbot. The overall goal is to provide trustworthy, medically acuate guidance in a secure and anonymous environment.

Because the system operates within a highly sensitive domain, it is crucial that anonymity and data security are prioritized throughout the design [18, 30]. The aim is to create a solution that users can trust, both technically and ethically, and to guarantee medical credibility through sources validated by Danish health organizations. This not only distinguishes the chatbot from typical LLM-based solutions but also ensures that it can contribute to reducing misinformation and promoting positive sexual-health behavior among young users.

4.1 Alternative Deployment Approaches

To achieve these goals, an initial review of possible technical approaches was carried out. Various chatbot solutions were identified and evaluated based on their advantages and disadvantages, particularly in relation to the system's key requirements: high data security, user anonymity, and control over the data used for model training and grounding [1, 12, 38]. The following present the most relevant categories of deployment options and how they align with, or conflict with, these requirements.

Cloud-based managed services. One initial option was to rely on fully managed LLM services such as OpenAI or Azure OpenAI [25, 34]. These platforms offer strong performance, high availability, and easy scaling. Their main advantage is immediate access to state-of-the-art models with minimal development and maintenance effort. However, every query and embedding request in such a setup would be processed externally, beyond the owner's physical control. Even if personal identifiers were removed, metadata and access logs could still make individual sessions traceable. In a sexual-health context, this introduces a significant privacy risk. It would also be impossible to ensure that the model's training data or moderation filters are aligned with the medically validated information required for this project.

Private cloud on Microsoft Azure. A second approach was to use a private deployment in Microsoft Azure configured under the EU Data Boundary [25]. This setup would allow sensitive data to remain within European regions and offers advanced security features such as managed identity, key vaults, and encryption-at-rest. Azure documentation also indicates that it may be possible to disable data logging on request [25]. While this approach would meet most compliance standards, it still depends on a third-party provider. As a result, it would not fully satisfy the project's goal of complete data ownership or allow full control over the sources that inform the model's responses.

On-premises deployment. The final approach was a fully on-premises setup hosted within Aalborg Universities secure infrastructure, or any similar institutional environment adopting the solution. This option requires more technical work and ongoing maintenance but provides total control over data flow, network security, and model configuration. All conversation data, embeddings, and reference material remain entirely local. It also allows full transparency regarding how prompts are processed and how models are updated or replaced [13]. In this setup it is possible to implement a Retrieval-Augmented Generation (RAG) technology, which enables the developer to control exactly which data the model uses as contextual grounding for its answers [13]. Although this solution demands greater effort in design, technology selection, development, and deployment, and would require maintenance in a production setting, it is the only option capable of meeting all three of the project's core requirements: security, anonymity, and medical credibility [24].

The three options represent a familiar balance between convenience and control. Cloud-based services offer scalability and ease of use but at the cost of data sovereignty. Private-cloud solutions improve compliance but still rely on external vendors. On-premises hosting requires greater development effort but provides complete data isolation and accountability.

Given the project's focus on anonymity and medical credibility, the on-premises setup was selected. It uses a Retrieval-Augmented Generation (RAG) structure to ensure that model outputs are grounded in validated medical information, eliminates third-party dependencies, and offers full transparency. The main trade-off is higher development and maintenance effort, which was considered acceptable given the ethical and privacy-sensitive nature of the domain [3, 26]). It should be noted that the model's ability to provide relevant and accurate answers must still be demonstrated through testing, an uncertainty that applies to all potential approaches.

4.2 System Overview

Figure 1 shows the resulting architecture. It consists of four main sections: *Client*, *Ingest*, *Core Chat System*, and *Ops*. Together these

components create a self-contained, auditable environment for secure dialogue and medically accurate knowledge retrieval.

Client layer. The client layer consists of a simple React-based web application that communicates with the system exclusively through a REST API. This choice allows rapid development of a functional user interface with minimal technical overhead, which is appropriate for a prototype intended for evaluation rather than production deployment. In contrast to directly integrating the chatbot into an existing platform such as the *Sex & Samfund* website, this standalone client provides full control over the interface and interaction flow. This reduces dependencies on external systems and minimizes technical and organizational constraints, allowing the project to focus on evaluating system behavior and user trust rather than integration complexity. In the current prototype, the reverse proxy has not yet been implemented, but this simplified configuration is sufficient for demonstration purposes. In a production environment, communication would take place through HTTPS connections managed by a reverse proxy. It would also be possible to configure the reverse proxy to ensure that no data is logged, which is particularly important in this context. The client could further include an embeddable widget for external websites such as *Sex & Samfund*, making the chatbot easily accessible without requiring a separate interface. By maintaining consistent API endpoints, the system can support multiple front-end contexts without exposing internal logic.

Reverse proxy. Although not implemented in the prototype, the reverse proxy is included in the architecture because it would be essential in a production setup. All external communication would pass through this component, which could be implemented using *Caddy* or *Traefik* [6, 46]. The proxy handles HTTPS termination, routing, and rate limiting. This isolates internal services from direct access and ensures that logs contain only metadata, not message content.

Core API and orchestration. The backend, built with *Java Spring Boot*, manages the orchestration of all components [43]. It handles incoming requests, performs validation, and coordinates calls to both the language model and the vector database. The API ensures that each user query follows a consistent processing pipeline. The REST API is designed to be stateless, meaning that no user sessions or chat histories are stored. This also implies that once a session ends, the conversation cannot be resumed, a deliberate trade-off made to guarantee that no user data is retained, thereby maintaining full anonymity and data privacy.

Retrieval and embeddings. The system uses a vector database (*Qdrant* [36]) to store semantically indexed passages of medically credible material. At present, the *Scheduler* is executed manually, but in a production setup it could run automatically to trigger the crawler at fixed intervals, ensuring that the database is updated whenever partner websites are revised. A custom *crawler and parser* written in Java extracts information from credible health sources. The data is then processed into smaller text chunks and embedded using a local *BGE* embedding model via *Ollama*. This setup allows continuous updates while maintaining full control over what content is indexed [33]. The indexing pipeline runs independently from the chat system to prevent interruptions. Only documents from trusted sources are included, ensuring that all answers are grounded in verified medical information. In some cases, it may be desirable to perform data cleaning at this stage, for instance, by filtering or excluding certain types of content. In the current setup, however, no filtering is applied - all content from the source pages is stored in the database exactly as it appears on the website.

LLM inference. The system supports local inference through *Ollama*, which manages open-weight models on the server. For the prototype, the model *Qwen3-4B-Instruct-2507-GGUF:Q4_K_M* were selected because it provide a good balance between performance, hardware efficiency, Danish-language capability, and instruction-following quality. The models can be swapped or combined within the same environment, allowing comparative testing under identical conditions. Running them locally ensures that no user input or context leaves the secure environment, and the models can also operate fully offline, which further strengthens data privacy and system independence [33].

Ops and monitoring. Operational components are included in the architecture but not implemented in the prototype. In a production system, these would handle encryption, key rotation, and system monitoring. Logs would exclude all chat payloads and store only minimal metadata for performance analysis. Backups and TLS secrets would be managed locally to ensure compliance with institutional data policies.

Based on this architectural setup, specific technologies were selected to operationalize the system in a way that supports the project's requirements for anonymity, data security, and medical credibility.

- **Docker** was used for containerization to simplify deployment and ensure consistent environments. It allows the full stack to be deployed on AAU servers with clear separation of services [7].
- **Qdrant** was chosen over alternatives such as Milvus or Weaviate because it offers good performance, easy configuration, and stability for medium-sized datasets [27, 36, 49].
- **Java Spring Boot** was used for the API layer as it integrates well with Java-based crawling tools and is a well-established framework for developing REST APIs. Most APIs are designed to interoperate easily with Java, which makes this a practical choice [43]. An additional convenience is that the system's developer has prior experience with this technology stack, which supports efficient implementation and maintenance.
- **Ollama** was selected for local LLM hosting because it provides a lightweight runtime for managing multiple open-weight models and allows quick reconfiguration without data leaving the system [33].
- **React** was chosen for the front end due to its component-based architecture and strong support for real-time, stateful interfaces. It allows code reuse between the standalone chat client and the potential embeddable widget for partner sites [37]. Alternatives such as Vue or Svelte were considered, but React was preferred for its maturity, ecosystem, and developer familiarity [45, 48].

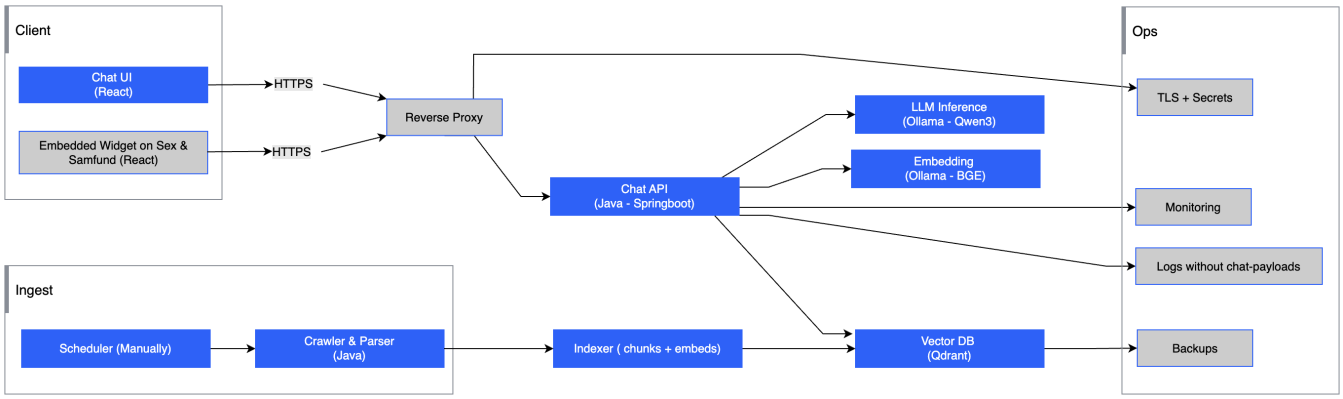

Figure 1: Overview of the final architecture. All data and models are hosted on-premises to preserve anonymity and data control. Grey elements are not implemented in the prototype for evaluation.

4.3 Model Selection and Complexity

Selecting an appropriate language model was one of the most technically demanding aspects of the project. It is not a single choice but a process involving evaluation of model families, size, context handling, and response characteristics. For this project, versions of *Mistral* and *Qwen3* were used as the primary models [33].

Why Mistral and Qwen3. *Mistral* offers efficient inference and strong instruction-following behavior, making it well suited for local deployment. *Qwen3* supports longer context windows, which is particularly important in a RAG setup where the model is continuously fed with contextual information. Another essential factor is that both models handles the Danish language. Using both models makes it possible to compare performance in terms of speed, accuracy, and conversational tone, allowing the system to be adjusted for different counseling scenarios [33].

Prior to the final selection of the model *Qwen3-4B-Instruct-2507-GGUF:Q4_K_M2*, a broader range of language models was evaluated, including models beyond the *Mistral* and *Qwen3* families. The Ollama models have different specifications, such as supported context length, GPU memory requirements, degree of optimization, and response behavior, including quality of response, and whether explicit "thinking" is exposed in the generated response. Model selection therefore involved a trade-off between competing priorities. Several of the evaluated models demonstrated response latencies exceeding 20 seconds, even when deployed on a high-performance GPU, which was considered unsuitable for an interactive counseling context. At the same time, it was essential that the model could generate high-quality, coherent responses without exposing internal reasoning processes in the final output. The selected *Qwen3* variant provided a balanced compromise between inference speed, response quality, and conversational clarity, making it appropriate for the intended deployment setting.

Selecting a model that performs reliably in a RAG configuration is crucial. The model itself is not trained further, instead, it receives contextual documents as part of the prompt, which it then uses to generate its answer. Model documentation typically includes information about which configurations the model performs best in, so it is important to review this carefully and, when possible, test several models in practice [3, 26]. In this project, this flexibility has been built into the system. From the user interface, it is possible to switch between models, and from a development perspective, additional models can be integrated relatively easily, provided they are compatible with the existing configuration. However, this model-switching feature was deliberately removed in the deployed prototype used for the trust evaluation. Exposing such functionality to participants was considered likely to confuse users and to introduce an additional variable that could influence trust perceptions.

Fine-tuning and configuration. After model selection, a range of parameters still influence the quality of the responses. These include chunk size, overlap, retrieval threshold. Small chunk sizes risk fragmenting meaning, while large chunks reduce precision. Similarly, lowering the retrieval score threshold increases recall but introduces irrelevant context. Adjusting these values requires iterative testing to find the right balance between accuracy and response time.

Data cleaning and validation. When crawling external websites, duplicates and boilerplate content can occur, such as identical information appearing under multiple URLs. Removing such redundancy is essential for reliable retrieval. The system therefore normalizes URLs, merges overlapping content, and filters unnecessary text.

4.4 Expert Evaluation

To complement the technical development and user-focused evaluation, expert feedback was incorporated as part of the system refinement process. The purpose of involving a domain expert was to assess the chatbot's medical accuracy, communicative clarity, and practical usefulness from a professional sexual-health perspective. While the pre-thesis included expert input at a conceptual level [24], this feedback focused specifically on the behavior of the implemented chatbot and the quality of its generated responses.

The expert consulted, from *Sex & Samfund*, had previously been responsible for the operation of the organization's home-testing services. This role provided extensive insight into the types of questions, misunderstandings, and information gaps that typically arise

among young users seeking sexual-health guidance. The expert was asked to freely explore the chatbot and provide general feedback, including factual verification, evaluation of dialogue quality, and reflections on tone, precision, and usability. In addition, the expert was encouraged to test the system using questions commonly encountered in professional counseling contexts, as well as scenarios known to cause confusion among users.

Overall, the feedback highlighted a need to improve linguistic quality, clarity, and depth of explanation across the chatbot's responses. These issues were described as recurring rather than isolated, suggesting systematic limitations rather than isolated errors. Several examples illustrated that responses could be grammatically imprecise, repetitive, or structured in ways that made key information difficult to identify. In particular, explanations of symptoms were sometimes perceived as unclear or poorly organized, making it challenging for users to distinguish primary symptoms from secondary information.

The expert also pointed out instances where medically relevant details were missing or insufficiently explained. For example, responses related to testing recommendations lacked important clarifications regarding different forms of unprotected sexual contact, and in some cases consequences of untreated infections were not adequately addressed for all genders. In addition, opportunities for preventive guidance - such as encouraging partner notification or explaining the rationale behind testing - were sometimes absent, even when highly relevant to the user's situation.

In more complex and sensitive scenarios, such as first-time sexual experiences involving unprotected intercourse, the expert noted that the chatbot demonstrated positive intentions through a supportive and empathetic tone. However, these responses also revealed significant weaknesses. Certain formulations were described as confusing, and explanations of pregnancy risk lacked a clear pedagogical structure. Notably, the feedback emphasized that responses did not sufficiently account for contextual factors such as menstrual cycles, nor did they consistently guide users toward appropriate next steps, such as seeking professional advice or involving trusted individuals in their support network.

Taken together, the expert feedback suggests that while the chatbot shows potential as a supportive entry point for sexual-health information, its current responses require refinement to meet professional standards for clarity, precision, and pedagogical quality. Importantly, several of the issues identified by the expert resemble challenges described in prior studies of health chatbots, where linguistic imprecision, repetitive responses, and limited contextual adaptation have been shown to affect the quality and reliability of generated advice [8, 18].

Linguistic inconsistencies, vague formulations, and repetitive phrasing reflect limitations in how models structure explanations, particularly when operating within constrained contextual grounding. Moreover, the tone and level of detail in responses are strongly influenced by prompt design and system-level instructions.

From a development perspective, the feedback indicates that improvements could be achieved without altering the overall architecture. Potential adjustments include refining the system prompt to better control tone, length, and structure of responses, as well as introducing clearer guidelines for summarization, prioritization of key medical facts, and avoidance of ambiguous phrasing.

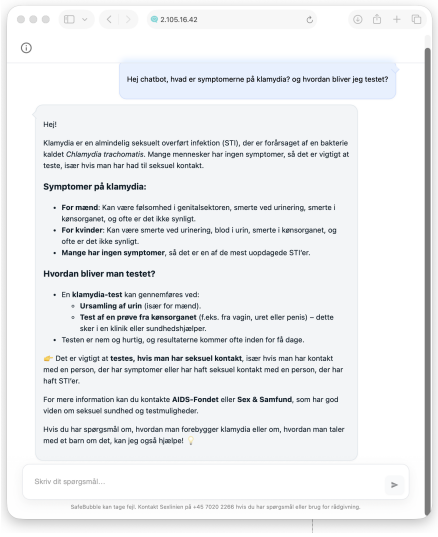

Figure 2: *SafeBubble*'s user interface

Importantly, the inclusion of expert feedback addresses a limitation that has been identified in prior research on digital sexual-health interventions. Reviews of existing STI-related applications have shown that many systems are developed and deployed with limited involvement of medical professionals, which can result in questionable medical accuracy [9]. This lack of systematic medical validation has been highlighted as a challenge for ensuring the quality and reliability of digital sexual-health information. By incorporating expert review of concrete chatbot responses, the present study address this limitation. The expert feedback provides both factual verification and domain-specific insight into how information is likely to be interpreted by young users seeking guidance. This makes it possible to identify shortcomings that are not easily detectable through technical testing alone, such as linguistic ambiguities, pedagogical gaps, or missing preventive recommendations.

4.5 The Final Prototype

The final prototype, named *SafeBubble*, is a fully on-premises, medically grounded chatbot application designed for evaluation with end users. A hosted version of the prototype was used during the evaluation and is available online[1]. The full implementation is available in a publicly accessible GitHub repository[2]. It is intended to provide trustworthy sexual-health guidance while preserving anonymity and strong data control. Figure 2 can be used to illustrate the user interface of *SafeBubble*, while Figure 3 can illustrate the end-to-end question-answer pipeline described below.

User-facing interaction. The user interface was deliberately designed to feel familiar and low-friction. Since the project evaluates user trust, the interface aims to minimize purely aesthetic influence so that trust judgments are driven primarily by the system's behavior, response, and perceived trust rather than visual styling. The interaction follows common chat conventions: the user submits a

[1] http://2.105.16.42
[2] https://github.com/FrederikkeFM/Thesis

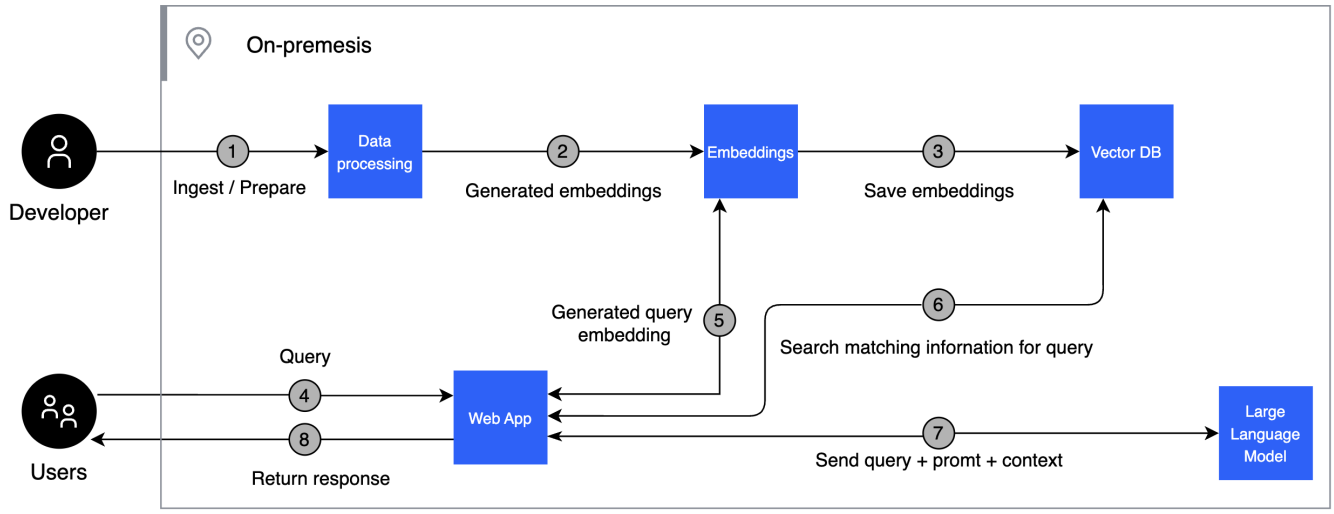

Figure 3: Overview of the end-to-end question-answer pipeline in *SafeBubble*.

question and receives a response presented as a message bubble in a conversational thread, similar to widely used chat platforms. This format supports sensitive information seeking by allowing users to phrase questions in their own words while maintaining a dialogue-like flow.

Medically grounded response generation. To ensure medical credibility, responses are generated using a Retrieval-Augmented Generation (RAG) approach grounded in pre-selected, medically validated Danish sources: *Sex & Samfund* and *Checkpoint (AIDS-Fondet)* websites [2, 40]. The content from these sources is collected through a custom crawler, after which the extracted text is segmented into smaller passages that can be retrieved as contextual support for the language model. Each passage is embedded and stored in a vector database, enabling semantic retrieval.

When a user submits a question, the system processes it through the same embedding procedure and retrieves the most relevant passages from the database based on similarity. In the current configuration, the system retrieves the five highest-scoring passages to serve as contextual grounding. These retrieved passages are then appended to the prompt, and the language model generates an answer based on the combined input. This pipeline aims to reduce hallucinations and align responses with accurate medical information, while still allowing questions to be asked in natural language.

Anonymity and data security by design. Because the system operates in a highly sensitive domain, anonymity and data security are treated as primary design constraints. The prototype does not log or store user messages, and it does not maintain user accounts or persistent sessions. All inference runs locally, and no requests are sent to external APIs, as for example OpenAI or Azure. A user question is used only transiently for generating the response and is not retained afterwards. This design supports the study's goal of ensuring that individual interactions are not traceable through stored conversation content.

Dialogue continuity without persistent storage. Although the system is intentionally stateless, the prototype supports a more dialogue-like interaction by including the model's immediately preceding response in each new prompt. This enables short-term conversational continuity, as follow-up questions and clarifications, without storing full chat histories. Since only the latest model output is carried forward, the amount of transmitted context remains small while still allowing the conversation to feel coherent. This approach represents a deliberate trade-off between conversational memory and strict anonymity.

Model behavior and language constraints. The prototype uses locally hosted, Ollama-compatible language models selected through preliminary testing to balance response quality, runtime performance, and Danish-language capability. An additional practical requirement was that the model should not output intermediate reasoning("thinking") as part of the visible response, as this can disrupt the intended counseling-like tone. Because the system operates fully offline, it does not search the internet or access external sources at runtime, responses are therefore constrained to the retrieved context and the model's underlying training data. The system prompt is configured to answer questions related to sexual-health in a natural conversational style. In the current prototype, system-prompts are written in Danish, and the system therefore generates responses in Danish, supporting other languages would require prompt adjustments and either asking for users preferred language or reading it from their browser configurations.

Modularity and future embedding contexts. Finally, the prototype is designed such that the conversational interface could be deployed in different front-end contexts. Beyond the standalone web client used in this project, the same functionality could be embedded as a widget in external platforms, for example as a pop-up element on *Sex & Samfund*'s website. This modularity supports the long-term vision of making medically validated information more accessible at the point where users already seek guidance.

5 MY STUDY

The methodological approach was designed to capture how participants form trust across three different digital systems. In line with established experimental practices in human-computer interaction, the study includes one target system and two comparison systems to enable meaningful interpretation of trust differences across conditions [19]. Rather than evaluating the developed system in isolation, this design allows trust to be assessed relative to existing and familiar alternatives. The *SafeBubble* chatbot served as the primary system under investigation.

To contextualize participants' trust evaluations, two comparison systems were included. *ChatGPT* was selected as a general-purpose conversational agent representing a widely adopted and recognizable chatbot technology with advanced dialogue capabilities. Prior research and expert input indicated skepticism toward simpler, rule-based chatbots commonly embedded in websites, particularly regarding their ability to support meaningful dialogue [24]. Including *ChatGPT* therefore enabled comparison with a state-of-the-art conversational system that many users already perceive as competent and useful. The *Sex & Samfund* website was included as a domain-specific comparison system representing an established and trusted source of sexual-health information in Denmark. Throughout the study, *Sex & Samfund*, together with *Checkpoint (AIDS-Fondet)*, has served as a key reference point, as both organizations are officially recognized by the Danish Health Authority. Including this system made it possible to examine how trust in an institutionally endorsed information source compares to trust in a newly developed chatbot grounded in the same medical domain. Together, the three systems represent distinct but relevant points of comparison: a novel, medically grounded chatbot, a highly familiar and widely adopted conversational AI, and an established domain authority. This configuration enables analysis of how different system characteristics, such as conversational form, domain specificity, and institutional credibility, shape distinct dimensions of trust.

The structure of the study follows established practices in human-AI trust research, where trust is examined through controlled exposure to a system, performance of domain-relevant tasks, and subsequent assessment using validated multi-dimensional trust scales. Similar methodological setups have been successfully applied in recent studies on trust in AI-based health and decision-support systems, demonstrating that such approaches are suitable for capturing meaningful differences in perceived trust across systems [16, 35]. The methods applied here therefore combine theoretical foundations from organizational trust [21], trust in automation [17], and multidimensional trust frameworks [22], together with controlled exposure to different digital systems and subsequent trust assessment using established questionnaire-based measures.

5.1 Participants

A total of 60 participants were recruited for the study, with 20 participants allocated to each system condition. Participants had a mean age of 23.5 years, and the sample included 33 women and 27 men. Participants were recruited through a combination of online channels. Initially, recruitment was conducted via a Facebook group dedicated to students at Aalborg University, commonly used for sharing surveys and recruiting participants for academic projects. However, after one week this approach resulted in limited participation. To reach the target sample size, recruitment was subsequently extended to the author's personal network by sharing the study on private social media platforms, including Facebook and Instagram. The author belongs to the target age group of the study and has an established network within this demographic, which facilitated broader participation. In addition, the study was shared in several public Facebook groups specifically intended for survey distribution and participant recruitment. As a result, participants were recruited from an uncontrolled and heterogeneous demographic background. Beyond age and gender, no additional demographic information was collected, as these variables were considered sufficient for contextualizing the sample in relation to the study's focus on young people's trust in digital sexual-health systems. The allocation was chosen to ensure sufficient statistical sensitivity for detecting differences between conditions. Earlier work has demonstrated that participant groups of approximately 20-30 individuals per condition are adequate for observing systematic differences in perceived trustworthiness, even in complex interaction environments involving multiple trust dimensions [16, 19, 35]. To contextualize the data and account for potential sensitivity in participant experiences, respondents were asked to report their age, gender, and whether they were currently experiencing symptoms for which they might consider seeking medical attention. The only restriction placed on participation in the study was that participants were required to be between 15 and 35 years of age. This age range was selected because young people represent the population group with the highest incidence of sexually transmitted infections, and because this demographic is a primary focus of national sexual-health initiatives led by the Danish Health Authority, as well as of counseling and test services and interventions provided by organizations such as *Sex & Samfund* and *Checkpoint* [24, 44]. Prior epidemiological data and practice-based insights indicate that individuals in this age group are both at elevated risk of STI transmission and more likely to seek digital and anonymous sources of sexual-health information. Limiting participation to this demographic therefore ensured that the evaluation reflected the characteristics, needs, and behavior of the intended users of the system [10, 24, 44].

5.2 Experimental Design

The project adopted a between-subjects design in which each participant interacted with one of three conditions: the *SafeBubble* chatbot, the *Sex & Samfund* official website, or *ChatGPT*. A round-robin allocation procedure was applied to distribute participants evenly across the three conditions and to minimize allocation bias. The choice of a between-subjects design reflects the understanding that trust is shaped by condition-specific characteristics and that direct comparison of multiple conditions within a single session may alter or contaminate trust formation. Each participant completed an identical set of tasks within their assigned condition and subsequently responded to a trust questionnaire adapted for the project. Condition-specific versions of the questionnaire were used to keep data clearly separated across conditions.

5.2.1 Tasks. The six tasks used in the project were developed through insights from experts interviews conducted in pre-thesis

work with advisors from *Sex & Samfund* and *Checkpoint*. These interviews highlighted three areas commonly encountered in sexual-health counseling: prevention, testing and treatment, and partner notification. The tasks were designed to reflect realistic scenarios within these three domains and to ensure that each condition was evaluated under conditions that mirror the questions and challenges users typically present in practice.

Participants were asked to imagine themselves in each scenario and use their assigned system to explore the relevant information. The scenarios involved identifying appropriate contraceptive options, seeking alternatives to in-person consultation following repeat chlamydia diagnoses, understanding chlamydia transmission and symptoms, locating testing options, finding guidance on how to formulate a partner-notification message, and investigating alternative ways of notifying partners when direct contact felt uncomfortable. All participants completed the same set of tasks across conditions. Each task was time-limited to two to three minutes, allowing participants to engage meaningfully with the system while maintaining comparability across conditions. This approach aligns with task-timing methods used in controlled human-AI trust evaluations, where limited interaction windows help standardize decision-making contexts and reduce the influence of external factors [17, 22].

5.2.2 Measures. Trust in the system was assessed using a structured questionnaire that combined adapted Ability, Benevolence, and Integrity (ABI) items with constructs from established trust-in-automation literature. The questionnaire was administered in Danish, while the item formulations were adapted from validated English-language instruments used in prior studies [17, 21, 22]. An English version of the questionnaire is provided in Table 1 for transparency and reference.

The ABI items were derived from the trustworthiness framework originally introduced by Mayer and Davis [21] and further operationalised and validated in later work by McKnight and Kacmar [22]. These dimensions capture perceptions of system competence (Ability), concern for user wellbeing (Benevolence), and adherence to consistent and ethical principles (Integrity). The adaptation involved rephrasing items to fit the context of a digital sexual-health information system while preserving the conceptual meaning of each dimension.

To complement these items, the questionnaire incorporated constructs identified in Körber's trust-in-automation framework [17], including Reliability and Competence, Understandability and Predictability, Familiarity and Experience, Propensity to Trust, Intention of Developers, and general Trust in Automation. These constructs were included to capture dimensions of trust that extend beyond trustee attributes and relate to system behavior, transparency, user expectations, and prior experience.

Participants rated their agreement with each statement on a seven-point Likert scale. The full questionnaire consisted of four Ability items, four Benevolence items, three Integrity items, nineteen items derived from trust-in-automation constructs, and a final item measuring behavioral intention to follow system recommendations. Responses were aggregated into composite trust dimensions by averaging items measuring the same underlying construct. This multidimensional measurement approach reflects contemporary human-AI trust research, where trust is analyzed at the construct level rather than through individual items, enabling systematic comparison across conditions [16, 35].

5.2.3 Procedure. Participants accessed a custom-built test website,[3] where assignment to system conditions was handled using a round-robin allocation method. Rather than relying on purely probabilistic randomization, this approach ensured an even distribution of participants across the three conditions while maintaining an unpredictable assignment order at the individual level. Participants completed the six tasks sequentially using the assigned condition. Upon completion of the tasks, participants provided demographic information, followed by completion of the full trust questionnaire. The order in which the questions were presented was randomized. This was done to minimize potential order effects and reduce the risk that earlier questions would influence participants' responses to later questions. Randomizing question order was particularly important given the conceptual overlap between questions, as repeated exposure to similar topics could influence how participants evaluated the condition across trust dimensions [17, 19, 22].The final stage involved rating their intention to follow any recommendations provided by the system. The entire procedure took approximately 15 to 20 minutes.

5.3 Data Analysis

The purpose of the data analysis was to examine whether perceived trust differed systematically across the three conditions, and to identify which dimensions of trust were most affected by the type of system participants interacted with.

For each participant, composite trust scores were calculated by averaging responses to questionnaire items associated with the same trust construct, as described in subsubsection 5.2.2. These aggregated measures served as the dependent variables in the subsequent statistical analyses.

To examine differences in trust between the three conditions, one-way analysis of variance (ANOVA) were conducted [19]. Each ANOVA compared the mean trust scores of participants across the three systems for a single trust dimension. The conditions (*SafeBubble* chatbot, *Sex & Samfund* website, and *ChatGPT*) served as the independent variable, while the trust dimension under consideration served as the dependent variable.

The use of one-way ANOVA is appropriate because each participant interacted with only one system and trust was measured using numerical scale scores (Likert scale). This analytical approach is consistent with methods used in recent studies examining trust formation in AI-based systems, where trust dimensions are compared across independent conditions [16, 35].

The resulting analyses indicate whether differences in perceived trust across systems are statistically detectable for specific trust dimensions, providing a basis for interpreting how system type influences distinct aspects of trust.

6 FINDINGS

This section presents the findings from a comparative, between-subjects evaluation of three conditions: *ChatGPT*, *Sex & Samfund*,

[3] http://130.225.39.218. The website will be taken offline in early February 2026.

and *SafeBubble*. The findings are based on questionnaire data measuring trust-related constructs, behavioral intention, and demographic characteristics. Results are reported descriptively and through one-way ANOVA in order to identify differences between conditions.

A total of 60 participants completed the questionnaire and were evenly distributed across the three system conditions. The questionnaire consisted of demographic questions and multiple trust-related constructs derived from established frameworks, including ABI (Ability, Benevolence, Integrity) and Trust in Automation (TiA), as well as a measure of behavioral intention.

For each construct, individual item responses were aggregated into composite scores per participant. These composite scores were then grouped by system condition and used as input for the statistical analysis. All statistical tests were conducted using IBM SPSS Statistics tool.

6.1 Demographic Characteristics

Demographic data were analyzed to assess the comparability of participant groups across system conditions. Figure 4 presents the distribution of age across participants, with a mean age of 23.5 years. The age distribution shows a comparable spread across all three system conditions, with no clear clustering or systematic skew towards a specific age range.

In total, the participant sample consisted of 33 women and 27 men, with this distribution being consistent across all three system conditions. This balanced gender distribution supports the comparability of the groups and reduces the likelihood that observed differences are driven by gender-related factors.

Participants were also asked whether they experienced problems for which they might require medical advice, as a proxy for the relevance of sexual-health information at the time of the evaluation. Across the full sample, 60% of participants reported not experiencing such problems, while 40% reported that they did. The distribution of responses was comparable across system conditions, indicating no systematic differences in prior medical support needs between groups.

Overall, the demographic characteristics suggest that the three system conditions are comparable, supporting the validity of subsequent comparisons between conditions.

6.2 Comparison of Trust-Related Constructs

Differences between system conditions were examined using one-way ANOVA for each construct. Homogeneity of variances was assessed using Levene's test as provided by SPSS. For most constructs, the assumption was met, however, violations were observed for Familiarity and Intention of Developers. These results are therefore reported with this limitation in mind. In addition, robust alternatives to the classical F-test (Welch and Brown-Forsythe tests) were examined as provided by SPSS to account for potential violations of homogeneity of variances. Normality was not explicitly assessed, however, the analysis relies on composite scores derived from multiple Likert-scale items and balanced group sizes across conditions, for which one-way ANOVA is generally considered robust.

The ANOVA results reveal statistically significant differences between system conditions for several constructs. Significant effects

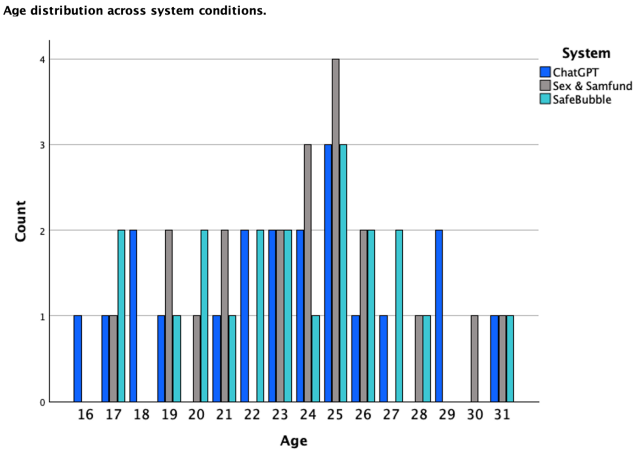

Figure 4: Age distribution across system conditions.

were found for Benevolence, $F(2, 57) = 13.31$, $p < .001$, Integrity, $F(2, 57) = 10.20$, $p < .001$, Reliability/Competence, $F(2, 57) = 7.86$, $p < .001$, Understandability/Predictability, $F(2, 57) = 9.05$, $p < .001$, Familiarity, $F(2, 57) = 12.47$, $p < .001$, and Intention of Developers, $F(2, 57) = 16.60$, $p < .001$. In contrast, no statistically significant differences were observed for Ability, $F(2, 57) = 2.19$, $p = .121$, Propensity to Trust, $F(2, 57) = 1.33$, $p = .272$, Trust in Automation, $F(2, 57) = 1.04$, $p = .361$, or behavioral intention, $F(2, 57) = 0.05$, $p = .952$.

Figure 5 illustrates the mean composite scores for each construct across system conditions. Overall, *SafeBubble* tends to score highest on constructs related to benevolence, integrity, and perceived developer intention, while *Sex & Samfund* generally scores higher than *ChatGPT* on several trust-related dimensions. *ChatGPT* scores relatively high on familiarity-related measures. These differences align with the significant effects observed in the ANOVA results.

6.3 Supplementary Analysis: Effects of System, Gender, and Context

To further examine whether the observed differences between system conditions were influenced by participant characteristics, a series of two-way ANOVAs were conducted for each construct with *System* and either *Gender* or *Self-reported need for medical advice* as between-subjects factors. The purpose of this analysis was to examine whether the observed differences between system conditions were consistent across participant groups, or whether evaluations differed depending on gender or participants' current need for medical advice. In other words, the analysis explored whether trust-related assessments of the systems changed as a function of these participant characteristics, beyond the overall system-level differences identified in the one-way ANOVAs [19].

Gender. Across constructs, the main effect of system remained consistent with the results of the one-way ANOVAs. No significant main effects of gender were observed for any of the constructs, indicating that overall trust evaluations did not differ systematically between female and male participants.

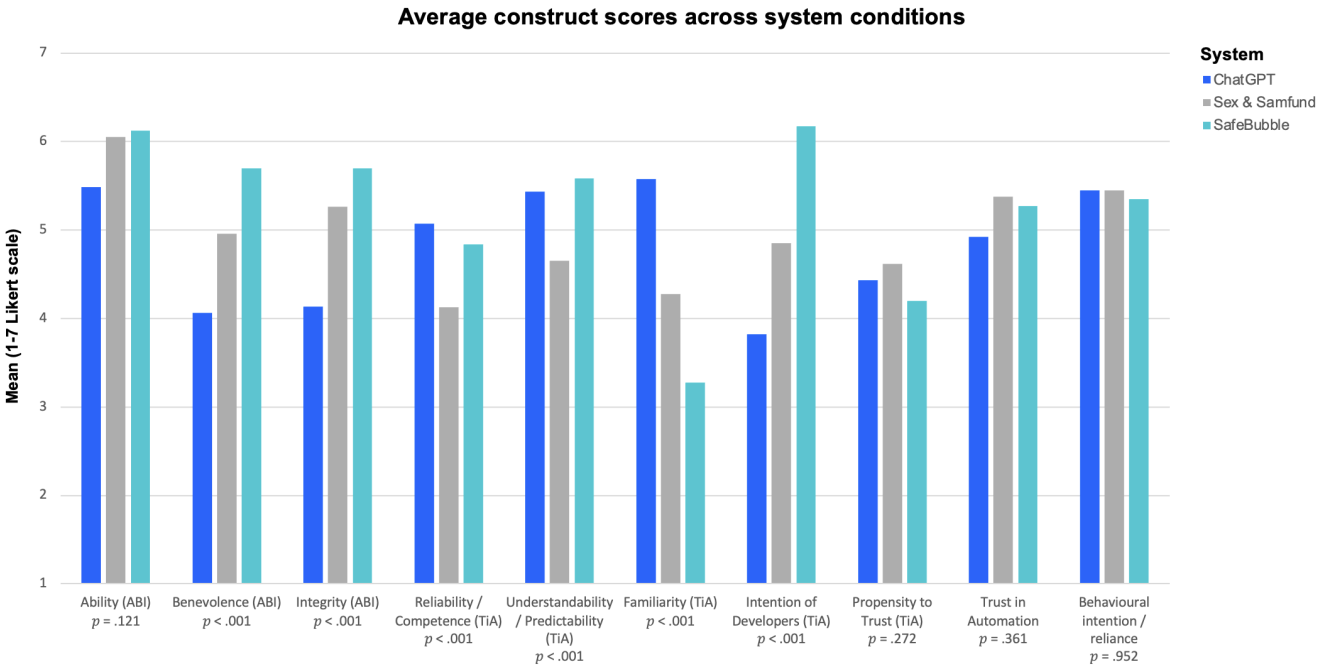

Figure 5: Mean composite scores for each construct across experimental conditions.

However, significant interaction effects between system condition and gender were identified for a limited subset of constructs. A significant System × Gender interaction was observed for Ability, $F(2, 54) = 3.68$, $p = .032$, Integrity, $F(2, 54) = 5.32$, $p = .008$, and Understandability/Predictability, $F(2, 54) = 4.21$, $p = .020$. These interaction effects suggest that the influence of system condition on these specific trust-related dimensions differed slightly between female and male participants. No significant interaction effects were found for the remaining constructs, including behavioral intention.

Self-reported need for medical advice. A parallel set of two-way ANOVAs was conducted with *System* and *Self-reported need for medical advice* (yes/no) as factors, in order to assess whether participants' current perceived relevance of medical information influenced trust evaluations. Across all constructs, the main effect of system mirrored the results of the one-way ANOVAs, while no significant main effects of medical advice were observed. Moreover, no significant interaction effects between system condition and self-reported need for medical advice were found for any construct.

This indicates that participants' trust-related evaluations were not systematically affected by whether they were currently experiencing issues for which they might seek medical advice. This result is consistent with the study design, which relied on scenario-based tasks rather than participants' own immediate information needs, and suggests that trust perceptions were shaped primarily by the systems themselves rather than by situational urgency or personal relevance at the time of evaluation.

Taken together, these supplementary analyses suggest that the observed differences between system conditions are largely robust across participant characteristics. While a small number of System × Gender interaction effects were identified for specific trust-related constructs, these effects were limited in scope and did not extend to behavioral intention or more general trust dispositions. No moderating effects were observed for self-reported need for medical advice.

Overall, system condition emerges as the dominant factor shaping trust-related evaluations, while participant characteristics such as gender or current medical relevance play a minimal role. As these analyses were exploratory and revealed only limited interaction effects, they are not treated as central explanatory factors in the interpretation of the findings, but rather as considerations for future work examining how individual differences may influence trust formation in greater depth.

6.4 Post-hoc Analysis of System Differences

While the one-way ANOVAs reported above establish whether statistically significant differences exist between system conditions for individual trust constructs, they do not indicate which specific systems differ from each other. Descriptively, Figure 5 illustrates the relative ordering of mean composite scores across conditions for each construct, suggesting systematic differences in how the three systems are evaluated. However, visual inspection alone is insufficient to determine whether these differences are statistically reliable. To further examine the nature of the observed effects, post-hoc multiple comparison tests were therefore conducted for all constructs [19].

Post-hoc tests were conducted to identify which specific system conditions differed from each other following significant one-way ANOVA results. Because this involves multiple pairwise comparisons, appropriate post-hoc procedures were used to ensure that

observed differences were not due to chance. Tukey's HSD was applied for constructs meeting the assumption of homogeneity of variances, while Games-Howell tests were used for constructs where this assumption was violated (Familiarity and Intention of Developers). Results are reported in terms of pairwise mean differences and homogeneous subsets, indicating which systems can be statistically distinguished.

Non-significant constructs. For Ability, Propensity to Trust, Trust in Automation, and behavioral intention, post-hoc analyses revealed no statistically significant pairwise differences between system conditions. These results are consistent with the non-significant one-way ANOVA findings for these constructs and indicate that none of the systems differed meaningfully from each other on these dimensions.

Significant constructs. For Benevolence, post-hoc comparisons revealed that *ChatGPT* was rated significantly lower than both *Sex & Samfund* ($p = .017$) and *SafeBubble* ($p < .001$). No statistically significant difference was observed between *Sex & Samfund* and *SafeBubble*. This suggests that the observed overall effect is primarily driven by lower benevolence ratings for *ChatGPT*.

For Integrity, post-hoc comparisons revealed that *ChatGPT* was rated significantly lower than both *Sex & Samfund* ($p = .007$) and *SafeBubble* ($p < .001$). No statistically significant difference was observed between *Sex & Samfund* and *SafeBubble*.

For Reliability/Competence, both *ChatGPT* ($p = .001$) and *SafeBubble* ($p = .016$) were rated significantly higher than *Sex & Samfund*, while no statistically significant difference was observed between *ChatGPT* and *SafeBubble*.

For Understandability/Predictability, *Sex & Samfund* was rated significantly lower than both *ChatGPT* ($p = .004$) and *SafeBubble* ($p < .001$). No statistically significant difference was observed between *ChatGPT* and *SafeBubble*.

For Familiarity, post-hoc results using the Games-Howell test showed that *ChatGPT* was rated significantly higher than both *SafeBubble* ($p < .001$) and *Sex & Samfund* ($p = .002$), while no significant difference was observed between *SafeBubble* and *Sex & Samfund*.

For Intention of Developers, Games-Howell comparisons revealed that *SafeBubble* was rated significantly higher than both *ChatGPT* ($p < .001$) and *Sex & Samfund* ($p = .004$). No statistically significant difference was found between *ChatGPT* and *Sex & Samfund*.

Overall pattern. Overall, the post-hoc analyses reveal a consistent yet dimension-specific pattern across the trust-related constructs. *ChatGPT* is evaluated significantly lower on benevolence and integrity, but scores highest on familiarity. *Sex & Samfund* is rated significantly lower on reliability/competence and understandability/predictability, while *SafeBubble* stands out by receiving the highest ratings on intention of developers and by performing at or above the level of the other systems across several constructs.

Importantly, these results reinforce the interpretation of trust as a multidimensional phenomenon, where differences emerge selectively across specific dimensions rather than uniformly across systems. Furthermore, the absence of pairwise differences for behavioral intention supports the conclusion that higher trust-related evaluations do not automatically translate into immediate intended behavior within the scope of this study.

6.5 Behavioral Intention

Despite significant differences across several trust-related constructs, no significant differences were found between system conditions for behavioral intention. Mean behavioral intention scores were comparable across all three conditions, suggesting that variations in perceived trust dimensions did not translate into measurable differences in intended behavior within the scope of this evaluation.

This finding indicates that while participants differentiated between systems on specific trust-related dimensions, such as benevolence, integrity, reliability/competence, and understandability/predictability, these differences did not result in divergent intentions to use or act upon the systems. Behavioral intention thus appears to be less sensitive to system-specific trust variations than anticipated, at least within the context and duration of the present evaluation.

One possible interpretation is that behavioral intention represents a more stable and context-dependent evaluative position than immediate trust perceptions formed during short-term system interaction. Prior research on digital sexual-health interventions suggests that intentions to act on health information are shaped not only by perceived system qualities, but also by factors such as personal relevance, situational urgency, prior experiences with similar services, and the perceived necessity of engaging with the information at a given moment [18, 29, 50]. In the present study, supplementary two-way analyses did not indicate that participants' self-reported current need for sexual-health guidance significantly affected trust evaluations. However, this should not be interpreted as conclusive evidence that personal or situational need has no influence on behavioral intention. Rather, it suggests that within the scope of brief, first-time interactions, immediate contextual need alone may be insufficient to shift stated intentions to act. Behavioral intention may therefore reflect broader motivational and contextual considerations that extend beyond the specific interaction evaluated in the study.

Importantly, the absence of statistically significant differences in behavioral intention across conditions does not contradict the observed differences in trust-related constructs. Instead, it aligns with prior work in trust and automation research, which has shown that perceived trustworthiness does not automatically translate into immediate reliance or action [17, 22]. Trust-related evaluations can be understood as a necessary precondition for engagement, but not a sufficient driver of behavioral commitment on their own. This distinction reinforces the relevance of treating trust as a multidimensional construct, where different dimensions may influence attitudes, perceptions, and behavioral outcomes in distinct ways, particularly in sensitive health-related contexts.

7 DISCUSSION

This section discusses the project's findings in relation to the research question of how, a sexual-health chatbot compares to other established systems, and to what extent, a sexual-health chatbot can be perceived as trustworthy by young people. The discussion

integrates three complementary perspectives. First, the quantitative findings are interpreted through established trust frameworks to examine how trust manifests across different dimensions and system conditions. Second, technical and architectural design choices are reflected upon in light of expert feedback, system constraints, and prior work on conversational agents in healthcare. Third, methodological considerations related to test setup, participant engagement, and measurement are discussed to contextualize the evaluation and clarify the conditions under which trust was assessed.

Together, these perspectives situate the results within existing HCI and trust research and support an understanding of trust in digital sexual-health interventions that accounts for empirical findings, system design, and evaluation context.

7.1 Discussion of Findings

The findings provide an empirical basis for discussing how, and to what extent, a sexual-health chatbot can be perceived as trustworthy by young people. Rather than indicating a single, uniform level of trust across systems, the results show that trust is distributed unevenly across multiple trust-related dimensions. This supports prior work in both HCI and trust research, which conceptualizes trust as a multidimensional and context-dependent phenomenon rather than a single aggregated construct [17, 20–22].

Across the three system conditions, statistically significant differences are observed for several system-specific trust dimensions, including whether the systems are perceived as acting in the user's interest (Benevolence), adhering to ethical principles and behaving responsibly (Integrity), providing consistent and dependable responses (Reliability/Competence), being easy to understand and predictable in their behavior (Understandability/Predictability), feeling recognizable or familiar to users (Familiarity), and being developed with care, responsibility, and appropriate intentions (Intention of Developers). In contrast, no significant differences are found for dimensions related to perceived capability to provide correct information (Ability), participants' general tendency to trust others (Propensity to Trust), their baseline trust in automated systems (Trust in Automation), or their stated intention to act on the information provided by the system (Behavioral Intention). This pattern suggests that perceived trustworthiness is shaped primarily by how users interpret the characteristics, intentions, and behavior of a specific system, rather than by stable individual trust dispositions. In relation to the research question, this indicates that a sexual-health chatbot can indeed be perceived as trustworthy, but that this trust is contingent on particular dimensions of trust rather than on general attitudes toward technology or automation.

The absence of significant differences in Propensity to Trust and Trust in Automation aligns with prior research describing these constructs as relatively stable user characteristics that are not easily influenced by short-term interaction [11, 20, 22]. Trust in automation research has consistently shown that baseline attitudes toward automated systems tend to persist unless users experience prolonged interaction or salient system failures [20]. The findings therefore indicate that the observed differences in perceived trustworthiness between systems cannot be explained by participants' general tendency to trust technology or automated systems. Instead, they appear to be driven by how the individual systems present themselves and behave during interaction, such as how responses are framed, how consistent and transparent the system appears, and how its purpose and intentions are communicated to users.

In contrast, dimensions such as Benevolence and Integrity appear highly sensitive to how the system is presented to users and to its perceived institutional grounding. These constructs reflect whether users believe that a system acts in their interest, follows ethical principles, and behaves consistently [21]. Prior research on digital health interventions highlights that trust is particularly influenced by perceived moral intent and legitimacy when systems operate in sensitive and potentially stigmatizing domains [29, 47, 50]. In the context of sexual-health, where users often seek reassurance and guidance under uncertainty, institutional grounding and perceived responsibility play a central role in shaping trust. The higher scores observed for these dimensions suggest that users' evaluations extend beyond technical functionality to include interpretations of why the system exists and whose interests it serves.

A notable finding is that while perceived Ability does not differ significantly across system conditions, Reliability/Competence does. This distinction highlights an important nuance in trust formation. Users may perceive multiple systems as knowledgeable or capable of providing correct information, yet differ in their confidence that these systems will behave consistently and predictably. Prior work in HCI and automation research emphasizes that predictability, transparency, and consistent system behavior are critical for sustained trust, often outweighing factual correctness when it is not accompanied by transparency, structure, or contextual guidance. [11, 17]. In sexual-health contexts, where users may prioritize clarity and reassurance over novelty or depth of information, reliability and understandability may therefore be more influential than perceived expertise alone.

Differences in Familiarity and Intention of Developers further underscore the interpretative nature of trust. Although these results should be interpreted with caution due to violations of statistical assumptions, the observed patterns align with prior findings showing that familiarity with a system or its underlying organization can reduce uncertainty and shape baseline expectations [20]. Perceived developer intention reflects users' assessments of whether a system is driven by care, responsibility, and public-health values rather than alternative motives. This dimension has been highlighted in prior qualitative studies of sexual-health chatbots, where trust is closely linked to perceptions of non-judgmental intent and professional responsibility [28, 30, 32]. The present findings suggest that such considerations remain salient even during brief, task-based interaction.

Despite significant differences across multiple trust-related dimensions, no corresponding differences are observed for behavioral intention. This dissociation between perceived trustworthiness and intended behavior has been reported in previous studies of digital health and AI-based interventions [30, 31, 50]. Behavioral intention in sexual-health contexts is influenced by a range of factors beyond trust alone, including stigma, perceived risk, personal readiness, and situational constraints [47]. In addition, participants' exposure to the systems in the present study was limited to a brief (approximately 10 minutes), first-time interaction, which may not have been sufficient to influence intentions related to behavior change.

The findings therefore suggest that while perceived trustworthiness may enable engagement with a sexual-health chatbot, it is not sufficient to produce immediate behavioral change within the scope of a short experimental interaction.

Taken together, the findings indicate that the perceived trustworthiness of a sexual-health chatbot is primarily shaped by relational and interpretative dimensions such as benevolence, integrity, reliability, and transparency, rather than by general trust attitudes or perceived technical ability. In relation to the research question, this suggests that a sexual-health chatbot can be perceived as trustworthy by young people, but that this trust is conditional and multidimensional. These results reinforce the importance of understanding trust in digital sexual-health interventions as a layered phenomenon, where institutional framing, perceived intent, and system behavior play a central role alongside technical correctness.

7.2 Technical Considerations and Expert Feedback

An important limitation of the present project concerns the timing and scope of expert involvement. Feedback from a medical expert from *Sex & Samfund* was received after the trust evaluation had already been conducted and could therefore not be incorporated into the system prior to user testing. As a result, the chatbot cannot be described as fully medically vetted in the strict sense, as this would have required iterative refinement of responses based on expert validation. This distinction is important, particularly in light of prior work emphasizing medical oversight as a key factor in establishing trust in digital health systems [18, 30].

The expert feedback primarily focused on the medical quality and phrasing of individual responses. While such feedback would be highly valuable in a setting where the language model could be retrained or fine-tuned on curated medical data, the present system architecture relies on a pre-trained model whose internal parameters cannot be modified by the developer. As a result, it is not possible to directly correct or override specific incorrect responses at the model level. Instead, the most actionable aspects of the expert feedback were those addressing general qualities such as tone, linguistic clarity, and response length. These aspects can be influenced through adjustments to the system prompt and retrieval configuration, allowing the developer to guide the model toward a more cautious, empathetic, and context-aware style of communication. This aligns with prior research showing that perceived trustworthiness in conversational agents is shaped not only by factual correctness, but also by how information is communicated [4, 28].

However, response quality emerged as a technical challenge in its own right. Although the selected language model explicitly supports Danish, testing revealed recurring issues with unnatural phrasing, incorrect translations, and the use of non-existing or semantically inaccurate Danish terms. This occurred despite the system providing Danish input throughout, including Danish context, user queries, and system prompts. The source of these issues appears to be the model's reliance on internal translation mechanisms when generating responses based on retrieved Danish text, resulting in degraded linguistic quality. Addressing this issue represents a substantial technical challenge and would likely require either the use of alternative models with stronger native Danish support, the integration of external API-based solutions, or model fine-tuning - each of which introduces new trade-offs related to cost, privacy, and data control. From a trust perspective, these findings underline that linguistic quality is not merely a usability concern, but a critical component of perceived credibility in sensitive domains such as sexual-health [32, 47].

7.2.1 Architectural Choices and Constraints. Although earlier expert interviews expressed skepticism toward chatbot-based solutions for sexual-health counseling [24], it is important to situate such concerns within the rapidly evolving landscape of AI-based conversational systems. The present system directly addresses prior concerns regarding overly simplistic or FAQ-like chatbots by employing dialogue-capable language models combined with retrieval-augmented generation. This enables more contextually relevant, conversational interactions that more closely resemble forms of digital counseling rather than static information delivery.

During the crawling and indexing process, it became evident that both *Sex & Samfund* and *Checkpoint* host extensive collections of medically validated content distributed across their websites. As AI-based chat interfaces increasingly replace traditional web search and manual navigation as primary entry points for information seeking, a conversational system has the potential to make this existing material more accessible to young people. In this sense, the chatbot does not replace existing counseling services, but rather functions as a mediating layer that can extend the reach of already validated educational resources [29, 50].

The resulting system architecture demonstrates that it is technically possible to build a sexual-health chatbot that meets strict privacy and data protection requirements. The on-premises deployment ensures that all data processing and model inference remain within institutional control, addressing concerns related to data ownership, compliance, and user anonymity that have been raised in prior work on digital health technologies [38]. The RAG-based design further ensures that responses are grounded in validated medical sources, thereby reducing, but not entirely eliminating, the risk of hallucinated or unverified information.

Compared with cloud-based solutions such as Azure OpenAI, the chosen architecture requires greater development effort and ongoing maintenance. However, this trade-off affords full transparency, data ownership, and easier verification of compliance, which are particularly relevant in sensitive health contexts. These design choices are consistent with the ethical and technical requirements outlined in the pre-thesis [24] and provide a realistic foundation for evaluating user trust in a system designed to prioritize privacy and anonymity through self-hosted, developer-controlled deployment.

Nevertheless, the system also reveals important limitations. Because the language model is trained on proprietary data that cannot be inspected or controlled by the developer, it cannot be guaranteed that responses rely exclusively on retrieved content from *Sex & Samfund* and *Checkpoint (AIDS-Fondet)*. Testing showed that the model occasionally introduced information not present in the provided context, indicating partial reliance on its underlying training data. This limitation reflects a broader challenge identified in recent

discussions of trust in AI systems, where transparency and control over model behavior remain incomplete despite architectural safeguards [1, 12].

The selection of source material for retrieval also proved critical. An example, is *Sexlinien* (a sub-page *Sex & Samfund*) that contains user-submitted questions and corresponding professional responses [41]. When content from this site was included in the retrieval database, the model sometimes produced misleading outputs by retrieving and reproducing non-factual user questions rather than medically grounded answers. In one case, a user query about potential chlamydia infection resulted in an affirmative response, reflecting the phrasing of prior user questions rather than appropriate medical guidance. This highlights the importance of careful selection and preparation of retrieval content, as not all institutionally hosted content is equally suitable for automated reuse. From a trust perspective, such errors can undermine perceived Ability, as they challenge users' confidence in the system's capacity to provide correct and medically reliable information, even when the underlying sources are associated with credible organizations [21, 22].

Taken together, these technical considerations illustrate how architectural and implementation choices directly shape the conditions under which trust can emerge. While the system demonstrates that a privacy-preserving, institutionally grounded sexual-health chatbot is possible, it also reveals that achieving medical vetting, linguistic quality, and reliable contextual grounding remains an ongoing challenge. These findings point toward future work focused on closer medical expert involvement to enable systematic validation of responses, improvements in linguistic quality to support more human-like and empathetic communication, and stronger control over the contextual material used during response generation. Together, these efforts should aim to move the system closer to cautious, dialog forms of guidance, thereby strengthening trust in digital sexual-health interventions.

7.3 Methodological Considerations

The methodological choices made in this project were shaped by the dual aim of enabling a meaningful comparison between different systems and evaluating perceived trustworthiness across multiple dimensions. While the chosen approach enabled a controlled comparison of system conditions, several methodological trade-offs and limitations should be acknowledged.

7.3.1 Test Setup, Participant Engagement, and Comparative Constraints. Several methodological considerations in this project relate to the overall design of the test setup, including task structure, questionnaire length, recruitment strategy, and cross-condition comparability. The evaluation combined scenario-based interaction tasks with a relatively extensive trust questionnaire in order to capture nuanced differences across multiple trust dimensions. While similar task-based approaches have been used in prior AI trust studies [16, 35], existing literature does not provide clear guidance to what extent participants should interact with a system relative to the length and depth of the following questionnaire. Existing studies vary considerably in experimental setup and level of participant supervision, ranging from in-person or guided evaluations to fully remote participation [29, 50].

In the present study, all participation was voluntary and conducted in an unsupervised home setting. This introduced a tension between ensuring sufficient exposure to each system and minimizing participant burden. Longer interaction periods increase the realism of the usage context by allowing participants to engage with the system in their own environment, but they may also discourage participation when recruitment relies on unpaid, voluntary engagement. This trade-off likely contributed to challenges in recruitment and completion rates and reflects a methodological issue in evaluating interactive health technologies outside controlled laboratory environments [19].

Related to this, the role and framing of the initial interaction tasks may not have been sufficiently explicit to participants. The tasks were intended to familiarize users with the assigned system and establish a comparable baseline for subsequent trust assessments. However, because participants were not required to explicitly report task outcomes, the purpose of the tasks may have appeared unclear. This issue was particularly salient for the static website condition, where information seeking inherently requires more effort compared to conversational interaction. While this may have increased cognitive load, it also highlights a meaningful contrast between traditional information navigation and AI-supported dialogue. From a methodological perspective, this illustrates the difficulty of designing tasks that are both system-agnostic and equally engaging across fundamentally different interaction paradigms, while at the same time underscoring the efficiency of conversational interfaces for information seeking [18, 28].

To enable direct statistical comparison across system conditions, the same questionnaire was used for all participants, as the questionnaire responses constituted the dependent variables in the ANOVA. While this ensured comparability, it also introduced limitations. Some questionnaire items were more naturally aligned with conversational systems than with a static website, which led to confusion for some participants. This reflects a known challenge in comparative HCI studies, where instruments developed to assess interactive systems may not transfer cleanly to non-interactive interfaces without adaptation [17]. Future studies could address this by retaining a shared set of trust-related items for statistical comparison, while supplementing these with additional condition-specific questions (for example dialogue quality for chatbots or navigation clarity for websites) analyzed separately to avoid compromising comparability.

Finally, recruitment challenges necessitated distributing the study across multiple platforms, reducing control over recruitment channels. While this resulted in a more diverse participant group within the intended age range, it also introduces potential sources of bias. In particular, participants who were aware of the study or the developer's involvement may have exhibited more favorable attitudes toward *SafeBubble*, potentially influencing trust-related constructs such as perceived benevolence or trust in developers. Although this does not invalidate the findings, it should be considered when interpreting differences between system conditions.

7.3.2 Content in SafeBubble. An additional methodological consideration concerns the framing of *SafeBubble* as a system grounded in content from *Sex & Samfund*. While transparency about data

sources is generally regarded as a trust-supporting design principle, this framing may have influenced participants' evaluations. If participants held prior attitudes, positive or negative, toward *Sex & Samfund*, these perceptions could directly affect their trust in *SafeBubble*'s content.

This introduces a potential confounding factor, as trust in the chatbot may partially reflect trust in the underlying organization rather than trust in the system itself. Such effects have been noted in prior trust research, where perceptions of institutional or organizational affiliation influence system evaluations independently of system behavior [11, 20]. While this effect is difficult to fully disentangle, acknowledging it is important for interpreting the findings.

7.3.3 Trust, Privacy, and Data Security. Finally, while privacy and anonymity were central design priorities in the technical implementation, these aspects were not directly operationalised in the questionnaire. Prior work has consistently identified data privacy and anonymity as critical factors for engagement with digital sexual-health services [24, 30, 32]. Although the questionnaire included constructs related to trust in developers and system intentions, these items may be interpreted in multiple ways, such as trust in content quality, ethical intent, or data handling practices, without explicitly isolating perceptions of data security.

Trust constructs most closely related to this aspect include integrity and benevolence [21, 22], as well as trust in automation constructs related to system dependability and reliability [17]. Incorporating more explicit measures of perceived data security, anonymity, and responsible data handling in future evaluations would make it possible to examine whether, and to what extent, these factors directly influence users' trust in sexual-health chatbots. This would allow a more fine-grained analysis of how different forms of trust contribute to system acceptance and adoption in sensitive health contexts.

Taken together, these methodological considerations highlight both the strengths and challenges of evaluating trust in AI-based sexual-health interventions. By adopting a comparative study approach using a controlled between-subjects design, combined with robust statistical analysis, the study provides a structured and credible basis for examining perceived trustworthiness across different system conditions. Within this setup, the results primarily reflect participants' trust perceptions formed through short, first-time interactions with the systems, rather than trust developed through long-term use or repeated engagement. This framing is important for understanding the scope of the findings and situates the results within the specific evaluation context used in this study.

7.4 Future Work

The findings and reflections presented in this study point toward two complementary directions for future work: further development of the *SafeBubble* prototype itself, and broader research directions for evaluating trust in sexual-health chatbots. These directions are outlined below.

Further development of the SafeBubble prototype. One direction for future work concerns completing and strengthening the *SafeBubble* system as a medically grounded sexual-health chatbot. A central aspect is closer and more systematic integration of medical expertise throughout the development process. While expert feedback in the present study provided valuable insights into tone, phrasing, and response quality, future iterations would benefit from iterative involvement of medical professionals. This would enable structured validation of responses and support the development of a system that can more confidently be described as medically vetted, in line with prior recommendations for AI-based health interventions [18, 30].

Beyond expert involvement, further development should also address technical factors that shape how medical knowledge is communicated and grounded in the system. Response quality emerged as a critical area for improvement, future iterations could therefore explore language models with stronger native Danish support, as well as strategies for achieving more consistent and human-like responses without relying on full model training. This should be complemented by tighter control over the retrieval process, including improved curation of source material and exclusion of content types unsuitable for automated reuse, in order to reduce the risk of misleading responses and strengthen users' confidence in the system's medical reliability and overall trustworthiness.

Future research directions. In addition to system-specific development, future work should also extend the empirical investigation of trust in sexual-health chatbots more broadly. While the present study assessed trust across multiple established dimensions, future studies could incorporate more explicit measures of perceived data security, anonymity, and responsible data handling. Such measures would allow for a more fine-grained examination of how privacy-related trust factors interact with other trust dimensions, as suggested by prior work [24, 32].

Finally, future research should examine trust development over longer periods of use. The present findings reflect trust perceptions formed during brief, first-time interactions. Longitudinal or repeated-use studies would make it possible to investigate how trust evolves over time, how it relates to continued engagement, and whether increased familiarity leads to changes in behavioral intention. Such work would provide a more comprehensive understanding of how trust in sexual-health chatbots is established, sustained, or challenged in real-world use contexts.

8 CONCLUSION

This study examined to what extent a sexual-health chatbot can be perceived as trustworthy by young people, and how such a system compares to other established alternatives. The findings demonstrate that a sexual-health chatbot can be perceived as trustworthy, but that trust is not uniform across systems or dimensions. Instead, young users differentiate between specific trust-related aspects, resulting in distinct trust profiles for each system.

Across the evaluated systems, *SafeBubble* was perceived as at least as ethically grounded and well-intentioned as *Sex & Samfund*, while being evaluated as more reliable, more understandable, and clearer in terms of interaction. This indicates that institutional authority and trusted content alone are insufficient to ensure positive interactional evaluations, and that how sexual-health information is delivered plays a central role in shaping perceived trustworthiness.

When compared to *ChatGPT*, *SafeBubble* was perceived as similarly usable and predictable, but as more ethically grounded and responsibly motivated in a sexual-health context. This indicates that while general-purpose conversational AI may feel familiar and easy to use, it does not elicit the same level of normative trust when evaluated in a sensitive health domain.

Taken together, the findings indicate that a sexual-health chatbot implemented as *SafeBubble* can achieve trust levels comparable to, and in several trust-related dimensions exceeding, those of established systems. In particular, *SafeBubble* was perceived as at least as ethically grounded and well-intentioned as *Sex & Samfund*, while being evaluated as more reliable, understandable, and interactionally clear. Furthermore, *SafeBubble* was perceived as similarly usable but more ethically grounded and responsibly motivated in a sexual-health context than *ChatGPT*, highlighting the value of a domain-specific chatbot over a general-purpose conversational AI.

Importantly, none of the systems elicited higher behavioral intention, underscoring that trust should be understood as an enabling condition rather than a direct driver of action. The findings of this study should be interpreted in light of short, first-time interactions and the absence of causal links between specific technical or design choices and trust outcomes.

Supplementary analyses further showed that participants' situational context, including whether they currently experienced a need for medical advice, did not significantly influence trust evaluations. This suggests that the observed differences reflect broader assessments of the systems themselves rather than participants' immediate personal needs.

Future work should therefore examine how trust develops through repeated interaction over time, incorporate more systematic medical expert validation of chatbot responses, and empirically explore how technical configurations influence trust-related outcomes. Overall, the findings indicate that chatbot-based interfaces such as *SafeBubble* can be trusted to communicate sexual-health content in a more reliable and understandable manner than traditional website-based solutions, and that achieving this level of trust requires a chatbot designed specifically for the sexual-health domain rather than a generic generative AI model.

REFERENCES

[1] Saleh Afroogh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. 2024. Trust in AI: progress, challenges, and future directions. *Humanit. Soc. Sci. Commun.* 11, 1568 (Nov. 2024).

[2] AIDS-Fondet. 2025. Seksuel sundhed for dig - Checkpoint. https://aidsfondet.dk/checkpoint

[3] Ivan Belcic. 2025. What is retrieval augmented generation (RAG)? https://www.ibm.com/think/topics/retrieval-augmented-generation

[4] Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. (2001), 396–403.

[5] Robert Bowman, Camille Nadal, Kellie Morrissey, Anja Thieme, and Gavin Doherty. 2023. Using Thematic Analysis in Healthcare HCI at CHI: A Scoping Review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18.

[6] Caddy. 2025. Caddy Documentation. https://caddyserver.com/docs/quick-starts/reverse-proxy

[7] Docker. 2025. Docker Docs. https://docs.docker.com

[8] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health* 4, 2 (June 2017).

[9] Jo Gibbs, Voula Gkatzidou, Laura Tickle, Sarah R. Manning, Tilna Tilakkumar, Kate Hone, Richard E. Ashcroft, Pam Sonnenberg, S. Tariq Sadiq, and Claudia S. Estcourt. 2017. 'Can you recommend any good STI apps?' A review of content, accuracy and comprehensiveness of current mobile medical applications for STIs and related genital infections. *Sex. Transm. Infect.* 93, 4 (June 2017), 240–246.

[10] Andreas Brandt Gormsen, Jon Erik Fraes Diernæs, Steen Hoffmann, and Uffe Koppelhus. 2018. Klamydia og lymphogranuloma venereum. *Ugeskrift for Læger* 180, 20 (May 2018), 889–893. https://pubmed.ncbi.nlm.nih.gov/29798754

[11] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors The Journal of the Human Factors and Ergonomics Society* 57, 3 (May 2015), 407–434.

[12] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John C Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. Position: TRUSTLLM: trustworthiness in large language models. 813 (2024), 105 pages.

[13] IMB. 2025. Retrieval Augmented Generation. https://www.ibm.com/architectures/hybrid/genai-rag

[14] Lena Jakob, Theresa Steeb, Zeno Fiocco, Teodora Pumnea, Sophia Nomi Jakob, Anja Wessely, Christoph Clemens Rothenberger, Titus Josef Brinker, Lars Einar French, Carola Berking, and Markus Vincent Heppt. 2020. Patient Perception of Mobile Phone Apps for the Care and Prevention of Sexually Transmitted Diseases: Cross-Sectional Study. *JMIR mHealth and uHealth* 8, 11 (Nov. 2020), 1–10.

[15] Inyeop Kim and Uichin Lee. 2024. Navigating User-System Gaps: Understanding User-Interactions in User-Centric Context-Aware Systems for Digital Well-being Intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15.

[16] Naja Kathrine Kollerup, Joel Wester, Mikael B. Skov, and Niels van Berkel. 2024. How Can I Signal You To Trust Me: Investigating AI Trust Signalling in Clinical Self-Assessments. (2024), 525–540.

[17] Moritz Körber. 2018. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. *SpringerLink* (Aug. 2018), 13–30.

[18] Liliana Laranjo, Adam G. Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y. S. Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *J. Am. Med. Inform. Assoc.* 25, 9 (Sept. 2018), 1248–1258.

[19] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction* (2nd edition. ed.). Morgan Kaufmann, Cambridge, Mass.

[20] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* 46, 1 (March 2004), 50–80.

[21] Roger Mayer and James Davis. 1999. The Effect of the Performance Appraisal System on Trust for Management: A Field Quasi-Experiment. *Journal of Applied Psychology* 84 (02 1999), 123–136.

[22] D. McKnight, Vivek Choudhury, and Charles ("Chuck") Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 13 (09 2002), 334–359.

[23] D. Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Manage. Inf. Syst.* 2, 2 (July 2011).

[24] Frederikke Filtenborg Michaelsen. 2025. Digital Interventions for STIs: Identifying Areas and Opportunities for Design.

[25] Microsoft. 2025. Azure OpenAI. https://azure.microsoft.com/en-us/products/ai-foundry/models/openai

[26] Microsoft. 2025. What is retrieval-augmented generation (RAG)? https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-retrieval-augmented-generation-rag

[27] Milvus. 2025. Welcome to Milvus Docs! https://milvus.io/docs

[28] Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Syst. Appl.* 129 (Sept. 2019), 56–67.

[29] Kathryn E. Muessig, Emily C. Pike, Sara LeGrand, and Lisa B. Hightow-Weidman. 2013. Mobile Phone Applications for the Care and Prevention of HIV and Other Sexually Transmitted Diseases: A Review. *J. Med. Internet Res.* 15, 1 (Jan. 2013).

[30] Tom Nadarzynski, Alexandria Lunt, Nicky Knights, Jake Bayley, and Carrie Llewellyn. 2023. "But can chatbots understand sex?" Attitudes towards artificial intelligence chatbots amongst sexual and reproductive health professionals: An exploratory mixed-methods study. *Int. J. STD AIDS* 34, 11 (June 2023), 809–816.

[31] Tom Nadarzynski, Oliver Miles, Aimee Cowie, and Damien Ridge. 2019. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *DIGITAL HEALTH* 5 (Jan. 2019).
[32] Tom Nadarzynski, Vannesa Puentes, Izabela Pawlak, Tania Mendes, and Damien Ridge. 2021. Barriers and facilitators to engagement with artificial intelligence (AI)-based chatbots for sexual and reproductive health advice: a qualitative analysis. *Sex. Health* 18, 5 (Nov. 2021), 385–393.
[33] Ollama. 2025. Ollama's documentation. https://docs.ollama.com
[34] OpenAI. 2025. OpenAI. https://openai.com/da-DK/api/
[35] Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proc. ACM Hum.-Comput. Interact.* 8 (Nov. 2024).
[36] Qdrant. 2025. Qdrant Documentation. https://qdrant.tech/documentation/
[37] React. 2025. Learn React. https://react.dev/learn
[38] John Mark Michael Rumbold and Barbara Pierscionek. 2017. The Effect of the General Data Protection Regulation on Medical Research. *J. Med. Internet Res.* 19, 2 (Feb. 2017), 1–6.
[39] Sex & Samfund. 2025. Danske unge slår igen egen rekord i klamydia - se hvor udbredt sygdommen er i din kommune. https://sexogsamfund.dk/om-os/nyheder/danske-unge-slaar-igen-egen-rekord-klamydia-se-hvor-udbredt-sygdommen-er-din-kommune
[40] Sex & Samfund. 2025. Sex & Samfund. https://sexogsamfund.dk
[41] Sex & Samfund. 2025. Sexlinien. https://www.sexlinien.dk
[42] Nicolas Scharowski, Sebastian A. C. Perrig, Nick von Felten, Lena Fanya Aeschbach, Klaus Opwis, Philipp Wintersberger, and Florian Brühlmann. 2025. To Trust or Distrust AI: A Questionnaire Validation Study. (2025), 361–374.
[43] Spring. 2025. Spring Boot. https://docs.spring.io/spring-boot/index.html
[44] Sundhedsstyrelsen. 2019. Klamydiatilfælde i 2012-2018 blandt 15-29 årige fordelt på kommuner. , 14 pages. https://www.sst.dk/-/media/Udgivelser/2019/Klamydiatilfaelde-pr-kommune-2012-18-testrate-og-postivrate-2018-28_11_19.ashx?la=da&hash=48FCD0B9C6B1051DF835A7948D22FC8F76E769C7
[45] Svelte. 2025. Svelte Documentation. https://svelte.dev/docs
[46] Traefik. 2025. What is Traefik? https://doc.traefik.io/traefik/
[47] Bettina Trettin, Tine Vestergaard, and Anette Stensgaard. 2015. Understanding young people's barriers to sexually transmitted disease screening and meeting their needs: a focus group study. *Journal of Nursing Education and Practice* 5, 6 (2015), 81–86.
[48] Vue. 2025. Vue introduction. https://vuejs.org/guide/introduction
[49] Weaviate. 2025. Weaviate Database. https://docs.weaviate.io/weaviate
[50] Laura Widman, Jacqueline Nesi, Kristyn Kamke, Sophia Choukas-Bradley, and J. L. Stewart. 2018. Technology-Based Interventions to Reduce Sexually Transmitted Infections and Unintended Pregnancy Among Youth. *Journal of adolescent health : official publication of the Society for Adolescent Medicine* 62, 6 (June 2018), 1–21.

Note: AI-based tools were used for linguistic refinement and text editing.

A APPENDIX

Item (Likert 1–7)	Trust dimension / construct
This system is very of performing its task of providing sexual-health information.	Ability (ABI)
This system has strong knowledge about the topics it gives guidance on.	Ability (ABI)
I feel confident in this system's ability to provide accurate information.	Ability (ABI)
This system is well qualified to support users with sexual-health questions.	Ability (ABI)
This system seems concerned about my wellbeing.	Benevolence (ABI)
My needs and concerns appear important to this system.	Benevolence (ABI)
This system would not knowingly provide guidance that could harm me.	Benevolence (ABI)
This system seems to look out for what is best for me.	Benevolence (ABI)
This system appears honest in the way it communicates.	Integrity (ABI)
I do not have to wonder whether this system will be consistent in what it says.	Integrity (ABI)
This system seems guided by fair and transparent principles.	Integrity (ABI)
The system is capable of interpreting situations correctly.	Reliability/Competence (TiA)
The system works reliably.	Reliability/Competence (TiA)
A system malfunction is likely.*	Reliability/Competence (TiA)
The system is capable of taking over complicated tasks.	Reliability/Competence (TiA)
The system might make sporadic errors.*	Reliability/Competence (TiA)
I am confident about the system's capabilities.	Reliability/Competence (TiA)
The system state was always clear to me.	Understandability/Predictability (TiA)
The system reacts unpredictably.*	Understandability/Predictability (TiA)
I was able to understand why things happened.	Understandability/Predictability (TiA)
It is difficult to identify what the system will do next.*	Understandability/Predictability (TiA)
I already know similar systems.	Familiarity (TiA)
I have already used similar systems.	Familiarity (TiA)
The developers are trustworthy.	Intention of Developers (TiA)
The developers take my well-being seriously.	Intention of Developers (TiA)
One should be careful with unfamiliar automated systems.*	Propensity to Trust (TiA)
I rather trust a system than I mistrust it.	Propensity to Trust (TiA)
Automated systems generally work well.	Propensity to Trust (TiA)
I trust the system.	Trust in Automation (TiA)
I can rely on the system.	Trust in Automation (TiA)
I follow the system's recommendations.	Behavioural intention / reliance

Table 1: Questionnaire items used to assess trust in the system, grouped by trust dimension/construct. Items marked with * are reverse-coded.