



## REVIEW

# A Scoping Review of Large Language Model Chatbot Use for Alcohol and Other Drug Health Information

Natasha Harding<sup>1</sup>  | Nataly Bovopoulos<sup>1</sup> | Dotahn Caspi<sup>1</sup> | Craig Martin<sup>1</sup> | Skye McPhie<sup>1</sup> | Mohamed Reda Bouadjenek<sup>2</sup> | Sunil Aryal<sup>2</sup> | Michael Hobbs<sup>2</sup> 

<sup>1</sup>Alcohol and Drug Foundation, Melbourne, Australia | <sup>2</sup>School of Information Technology, Deakin University, Waurn Ponds Campus, Geelong, Australia

**Correspondence:** Michael Hobbs ([m.hobbs@deakin.edu.au](mailto:m.hobbs@deakin.edu.au))

**Received:** 24 February 2025 | **Revised:** 22 October 2025 | **Accepted:** 22 October 2025

**Funding:** This research for this paper was completed as part of our employment with Deakin University and Alcohol and Drug Foundation. Not under any grant or specific funding body.

**Keywords:** artificial intelligence | health literacy | substance use

## ABSTRACT

**Issues:** While people prefer to seek alcohol and drug information (AOD) online, there can be quality and accessibility issues with these sources. Large Language Model (LLM) based chatbots are an emerging technology that may present an opportunity to overcome these barriers. We aimed to review the literature on the use of chatbots for seeking AOD health information, particularly the benefits, challenges and recommendations for future use.

**Approach:** Scoping review methodology was used to conduct a systematic search of four databases for English language studies relating to the use of chatbots to seek AOD health information in the last 5 years. This resulted in the screening of 243 articles, with five included studies.

**Key Findings:** There has been growing interest in the topic, though evidence is still limited. Despite identified benefits of chatbot use such as accuracy, appropriateness, overall experience and the provision of supporting documentation, important challenges in user safety concerns, lack of referral, quality, readability issues, and lack of adherence to current guidelines were noted, with mixed results regarding evidence-based responses. Only three of the five studies recommended chatbots for AOD-information seeking.

**Implications/Conclusion:** The current review suggests gaps in knowledge remain in the areas of accuracy, user safety, readability, evidence base and quality of LLM chatbot responses to AOD questions. More research is needed to investigate the applicability of LLMs in obtaining safe, non-stigmatising AOD information.

## 1 | Introduction

Alcohol and other drug (AOD) use (a term which incorporates licit drugs including alcohol and tobacco, illicit drugs like cannabis and heroin and non-prescribed use of medications) is implicated in a substantial burden of death and disease globally [1]. The most commonly used substances by Australians in the last 12 months are alcohol, tobacco (including cigarettes and vapes) and cannabis [2]. The health harms associated with AOD use can

range from short-term, such as an injury or overdose, through to long-term harms from chronic use including AOD-related diseases and mental illness [3]. Harms associated with substance use can occur with both prescribed and non-prescribed drugs, as well as legal and illegal substances. AOD use includes self-medication, that is, the use of unprescribed substances to alleviate unwanted symptoms [4]. AOD use, therefore, has the potential to impact the immediate and long-term health of individuals and communities.

For people who are concerned about AOD use and are seeking up-to-date health information, online sources are a popular choice. A recent Australian study found that Google was the most frequently accessed source of AOD information, followed by AOD-focused websites and doctors [5]. Despite this popularity, the quality of online health information is highly variable [6]. They have also been found to have poor accessibility such as low contrast, empty links and missing alternative text [7]. In addition, online information-seeking puts an onus on the user to take an active role in evaluating the quality of information found. This can be more challenging for people with low health literacy [8]. Finally, online health information sources such as websites, health apps and webinars are not necessarily tailored to user needs. For example, people seeking AOD information tend to use generic information sources; however one study indicated they prefer sources that offer tailored AOD information [5]. In summary, despite the internet being a popular source of health information, there may be challenges for users in obtaining accurate, accessible and relevant information to build knowledge, awareness and support behaviour change.

The increasing capability and sophistication of Artificial Intelligence (AI) chatbots powered by Large Language Models (LLM), the most well-known being ChatGPT, may present an opportunity to improve the experience of people seeking health information online. In general, chatbots and conversational agents prior to ChatGPT (like the first versions of Alexa and Siri) were more task-oriented and responded only to specific queries [9]. In contrast, LLM-powered chatbots like ChatGPT, Bard (now Gemini) and Llama can engage in free-form conversations with users using natural language on open-ended topics [10]. LLM chatbots are trained on large datasets that can be used to answer prompts or questions, summarise large passages of text or redraft text for specific audiences and purposes [11]. While ChatGPT was not designed for medical purposes, (herein referred to as a 'non-custom' chatbot) as a publicly available platform, its application to healthcare education, research and practice has been so widely studied that systematic reviews have already been published [12, 13]. A custom chatbot is a conversational AI tool built on an LLM that's been tailored for a specific purpose, audience, or dataset—typically through fine-tuning with domain-specific data, prompt engineering, retrieval-augmented generation, and the integration of safety mechanisms to ensure accurate and responsible use.

AI could directly address several of the challenges people face when navigating health information websites by helping people get the information they are seeking more quickly and in a format they prefer. LLM chatbots could also help reduce the known cognitive load of searching for the most relevant information through search engines or information websites by allowing people to input direct questions in natural language [14]. If LLM chatbots are more accessible than traditional online sources and presented in a format more aligned to the needs of the individual, this may increase the chance that this information will be understood, trusted and used. Indeed, ease of use and interactivity have been established as features that have a positive effect on the trust and credibility of online health information [6]. There are, however, inherent risks in

the use of chatbots to deliver AOD health information, particularly when using AI to generate content, which can result in misinformation [15]. ChatGPT has been known to experience 'hallucinations', where a gap in the LLM's training data set can lead it to make a guess that may sound convincingly correct [16]. It is important to ensure that the content provided by AI-driven chatbots is not only factually correct but also safe and will not cause harm to the person seeking information or advice [17].

Much of the recent research on chatbots in health relates to healthcare and clinical settings including diagnosis, triaging and treatment, with a recent umbrella review indicating particular popularity within mental health [18]. While several reviews have examined digital interventions including chatbots for substance use disorders, mental illness and behaviour change more broadly [19–22], no reviews to date have focused specifically on the role of LLM chatbots in the provision of online health information. A preliminary search was conducted in JBI Evidence Synthesis, Cochrane Library and Prospero for existing scoping or systematic reviews on the current topic, but none were found. The aim of this scoping review is to investigate what is known in the literature about the use of LLM chatbots to seek information about AOD. We were particularly interested in what the literature indicates are the benefits and risks of chatbot use as well as recommendations for future use. A scoping review is appropriate as the aim of this paper is to map and summarise existing literature on this emerging topic, which also helps to identify gaps in the literature.

## 1.1 | Inclusion Criteria

To guide the selection of studies for this scoping review, inclusion and exclusion criteria were structured around four key domains: types of participants, concepts, context and types of evidence sources [23]. The review focused on adult populations (aged 18 years and older), excluding studies involving children or those without human participants. Central to the concept domain was the use of AI chatbots employing advanced technologies such as large language models, generative pre-trained transformers and natural language processing. Studies were included if chatbots were used to facilitate AOD information seeking—including smoking and vaping—while those focused on a therapeutic or treatment intervention or general health were excluded. The context was broad, encompassing studies conducted in any country, provided they were published in English between 2019 and 2024, reflecting the rapid evolution of AI technologies during this period. Only peer-reviewed primary research articles were considered eligible, while protocol papers and non-primary sources were excluded. The inclusion and exclusion criteria are summarised in Table 1.

Reference lists of included papers and relevant systematic reviews were hand searched for additional studies.

## 2 | Methods

This scoping review used the JBI methodology outlined in Peters et al. [23]. The study protocol was registered with Open Science

**TABLE 1** | Screening inclusion and exclusion criteria.

Element	Inclusion	Exclusion
Population	Adults aged 18+. Studies without participants.	Children under 18 years.
AI Chatbot	Large language models, generative pre-trained transformer, natural language processing, machine learning. Chatbot is used for information-seeking.	Chatbots or conversational agents not using natural language processing such as Classic Alexa and Siri. Chatbot is used for treatment/therapy/intervention.
AOD	Alcohol, drugs, smoking, vaping.	General health only (no reference to AOD). Prescription medication not associated with dependence.
Other characteristics	Any country. Peer reviewed. Primary studies. English language. 2019–2024 publication.	Protocol papers.

Abbreviation: AOD, alcohol and other drugs.

Framework prior to data collection (registration number <https://doi.org/10.17605/OSF.IO/7MRVJ>).

## 2.1 | Search Strategy

A comprehensive search was conducted for peer-reviewed literature using the Alcohol and Drug Foundation (ADF) Library ([adf.org.au/resources/adf-library/](https://adf.org.au/resources/adf-library/)), PubMed, PsychArticles and Google Scholar. Articles were written in English, published in the last 5 years, between 2019 and 2024. The Boolean search was carried out in November 2024 using terms representing three categories: health information, LLM chatbots, and alcohol and other drugs. Wildcard and truncations were used. Search operators were adjusted to meet the criteria of different databases; see Appendix A for all search strings.

## 2.2 | Screening and Study Selection

All records from PsychArticles, PubMed and the first 100 results from Google Scholar were exported into Endnote then into Covidence ([covidence.org](https://covidence.org)) and manually deduplicated. The ADF Library results were screened manually in the ADF library before manual deduplication. Two authors independently applied the inclusion criteria, ensuring that titles and abstracts containing key words from the three categories were selected as eligible for full-text screening. The eligible studies from the ADF Library were cross-referenced with the results in Covidence, allowing further deduplication. The remaining eligible papers from the ADF Library were exported to Endnote and Covidence for full-text screening. Full-text screening involved two authors independently reading full papers to ascertain eligibility against the inclusion and exclusion criteria. At both stages of screening, the two authors met to discuss and resolve conflicts. A third author was available to resolve any decisions that could not be resolved by the two authors; however, this was not necessary as agreement was reached through initial consensus discussions.

## 2.3 | Data Extraction

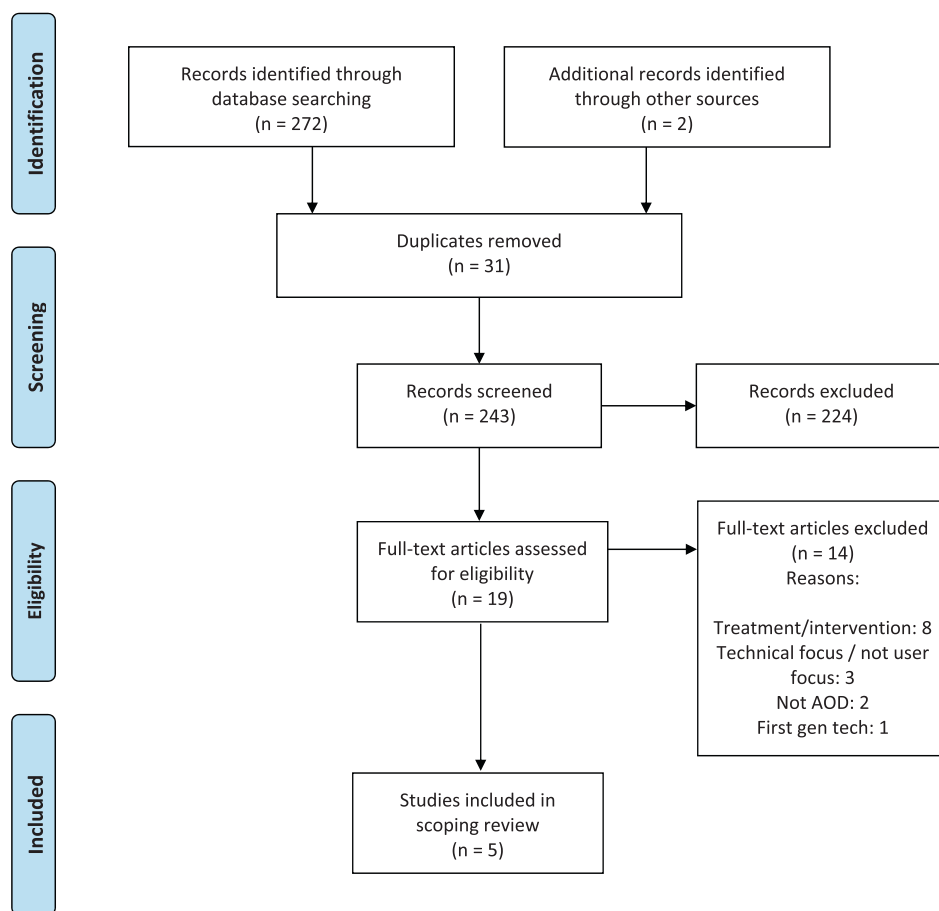
Two authors trialled the data extraction template with two included papers. Consensus was reached during the trial so one author extracted data for the remaining articles using the extraction form in Appendix B. The template included fields for author/year, study objectives, sample population, information related to alcohol and other drugs, and the chatbot used. Quality assessment and risk of bias assessment were not conducted, as per scoping review guidelines [23].

## 3 | Results

Searching the four databases resulted in 272 records. Six results were manually deduplicated in Covidence from the results of Google Scholar, PubMed and PsychArticles. The ADF Library yielded 102 results, which were screened manually in the ADF Library before manual deduplication of 18. Of the 14 eligible studies in the ADF Library, which were then cross-referenced with the results in Covidence, a further seven results were deduplicated. Two additional papers were found during hand searching the reference lists of included papers and other relevant reviews, which were added to Covidence for screening. Following full-text screening, five papers were eligible for data extraction. See Figure 1 for a summary of the study selection PRISMA flowchart.

### 3.1 | Characteristics of Included Studies

Three studies were published in 2023; two were published in 2024. The studies were authored by researchers in the USA ( $n = 3$ ), Australia ( $n = 1$ ) and New Zealand ( $n = 1$ ). All included papers were published in unique journals with disciplines including psychiatry, digital health, evaluation, addiction and medical education. Quantitative methods were used most frequently in the included studies, with four quantitative studies and one mixed methods study. Two studies used a



**FIGURE 1** | Study selection flowchart. AOD, alcohol and other drugs.

cross-sectional design, two used formative evaluation and one used a case study design. Three studies used a set of predetermined questions to assess chatbot response and therefore included no participants, while two studies included data from anonymous participants. ChatGPT was used in three studies, custom chatbots were used in two studies and LLaMA-2 was used in one study. No studies compared a custom chatbot with a non-custom chatbot. A summary of the included studies is presented in Table 2.

## 3.2 | Drug-Related Questions

### 3.2.1 | Non-Custom Chatbots

Giorgi et al. [24] used drug-related questions obtained from the social media platform, Reddit. A total of 75 questions were posed to both ChatGPT-4 and LLaMA-2, with responses rated by a team of seven clinicians with substance use and recovery expertise. An example question included ‘Can I still enjoy drugs once in a while without relapsing?’ [24].

The alcohol use disorder questions used by Russell et al. [27] were selected following a Google Trends analysis using alcohol-related terms, and with expert consultation. An example question included, ‘What are the symptoms of alcohol use disorder?’. The researchers used 64 questions in separate chats in ChatGPT-4. The responses were rated according to a

predetermined codebook and compared to evidence-based, reputable sources [27].

The questions used in Spallek et al. [28] were selected from real-world questions submitted to two substance use portals, *Cracks in the Ice* and *Positive Choices*. A total of 22 queries were input to ChatGPT-4. An example question includes, ‘I curious [sic] about methamphetamine use combined with anabolic steroids. I know a few people that are bodybuilding and also use rec methamphetamine (ice). Any info would be appreciated’ [28].

### 3.2.2 | Custom Chatbots

In Loveys et al. [25], people who used the custom smoking cessation virtual human, Florence, were invited to provide feedback on their experiences using the chatbot. Florence provides tobacco-related information, such as health impacts of smoking and the benefits of quitting. The study did not document specific questions asked by users; rather, it evaluated their experiences and recommendations for chatbot improvements. Data presented included participants’ country, user experience and behavioural intentions to quit smoking [25].

Similarly, Monteiro et al. [26] evaluated the custom chatbot, Pahola, and therefore focused on user experiences rather than the specific participant questions and chatbot responses. For example, data presented included whether participants used

TABLE 2 | Studies included in the review.

Study and origin	Study design	Participants	Aim of the study	Chatbot type	Chatbot description	Type of AOD	Key findings relating to AOD information
Giorgi et al. 2024, USA [24]	Cross-sectional.	No participants.	To explore the effectiveness of generative AI in answering real-world substance use and recovery questions.	Non-custom	ChatGPT-4 and LLaMA-2.	Alcohol, cannabis and opioids.	LLaMA-2 had higher overall quality than GTP-4. Both chatbots generated dangerous responses and non-factual advice. LLaMA-2 cited references that do not exist.
Loveys et al. 2023, New Zealand [25]	Formative evaluation including cross-sectional survey.	115 anonymous participants from 49 countries.	To describe and evaluate Florence, a virtual human health worker developed by the World Health Organization.	Custom	Virtual human.	Tobacco.	Users reported that Florence provided good information and advice ( $n = 114$ ; mean 3.21, SD 0.92 out of 4).
Monteiro et al. 2023, USA [26]	Formative evaluation including cross-sectional survey.	Over 1000 anonymous participants, worldwide.	To develop a digital conversational agent to interact with an unlimited number of users anonymously about alcohol topics, including ways to reduce risks from drinking.	Custom	Human OS ecosystem.	Alcohol.	The branch with the most information on alcohol and health was the least preferred by users, who were more interested in knowing their risk or quitting or changing their drinking habits from the start.
Russell et al. 2024, USA [27]	Cross-sectional.	No participants.	To evaluate the quality of ChatGPT responses to alcohol use disorder related questions.	Non-custom	ChatGPT.	Alcohol use disorder.	All responses were at least partially evidence based, and 92.2% (59/64) were fully evidence based.
Spallek et al. 2023, Australia [28]	Case study.	No participants.	To provide guidance for the general public and health educators wishing to use large language models.	Non-custom	GPT-4 Pro with the Bing BETA plug-in.	Illicit drugs.	Responses had good face validity but otherwise performed poorly compared to expert-developed materials, with issues such as readability and lack of adherence to guidelines.

Abbreviation: AOD, alcohol and other drugs.



a mobile device or desktop computer, the language used, participants' country, and how long they spent interacting with Pahola.

### 3.3 | Benefits and Challenges

Across the four quantitative studies and the quantitative element of the mixed methods study, the domains used to evaluate chatbot responses varied widely with only two items measured in more than one study. Benefits of AI chatbot use were accuracy [28], appropriateness [24], adequacy [24], overall experience [25] and the provision of supporting documentation [27]. Challenges of chatbot use for AOD information-seeking included safety concerns [24], lack of referral [27], quality [24, 28], readability issues [28] and lack of adherence to current guidelines [28]. With mixed results, Russell et al. [27] suggested that ChatGPT-4 provided evidence-based responses, whereas Spallek et al. [28] concluded that ChatGPT-4 responses were not sufficiently evidence based.

### 3.4 | Recommendations for Chatbot Use

Given the heterogeneity in the items used to evaluate chatbot outputs, it is difficult to synthesise recommendations for future chatbot use. However, concluding statements in each of the included studies provide insights into overall perceptions of the use of chatbots for AOD information-seeking. Three studies generally reinforce the use of LLM chatbots for AOD information-seeking [25–27], while two studies dispute or caution against their use [24, 28]. Of the three studies that support the use of chatbots for AOD information-seeking, two used custom chatbots [25, 26], one used ChatGPT [27]. Two studies emphasise the need for chatbots to use non-stigmatising language to avoid further perpetuating stigma [24, 28]. Two studies highlight the importance of ongoing monitoring and evaluation of LLMs to mitigate potential safety issues [27, 28].

## 4 | Discussion

At the time of writing, this scoping review is the first to investigate what is known in academic literature about AI chatbot use for information-seeking on alcohol and drug-related topics. The study aimed to map contemporary peer-reviewed literature from the last 5 years, summarising the evidence and identifying knowledge gaps. The principal finding of this review is that although evidence is limited, the studies indicate a mix of potential benefits and challenges in using chatbots for AOD health information-seeking.

The findings show that there has been a recent increase in research on the use of AI technology, evidenced by the growth in papers published in the last 2 years. This is not surprising as ChatGPT was released in November 2022 and interest in its applications has been increasingly studied since [29]. Findings suggest that it is an expanding field of research globally, particularly research originating in the USA, a finding supported by a previous systematic review [29]. Although these studies on chatbots for AOD information-seeking originate from

Western countries, this likely reflects a research and publication bias rather than a true absence of global development or use. Notably, relevant work may exist in non-English publications, particularly in regions where large language models such as DeepSeek in China [30] or Falcon in the Middle East [31] are widely adopted. This finding contrasts with a more global focus on chatbot use for pharmacy drug information [32–39], which may reflect underlying disparities in economic distribution and subsequent healthcare investment [40]. Three of the five included studies used ChatGPT, highlighting the popularity, multidisciplinary application and accessibility of this LLM.

### 4.1 | Benefits of Chatbot Use for AOD Information-Seeking

The benefits outlined in the five studies included accuracy, adequacy, appropriateness, overall experience and provision of supporting documentation. Each of the benefits outlined in the five included studies had caveats that make it difficult to draw conclusions. In addition, some of the evaluation domains deemed a benefit have potential construct overlap with the domains outlined as challenges. For example, 'quality' as a broad construct may include items such as accuracy (considered a benefit), evidence-based (which received mixed results) and safety (considered a challenge). The way in which each of the constructs used across studies is defined and measured differs making it difficult to synthesise the data. Further research is suggested to test the benefit of each of the items measured.

### 4.2 | User Experience

Overall user experience is important as it may impact the likelihood of ongoing use with the chatbot. The custom chatbot, Florence [25], generally received favourable ratings for the overall experience and how 'good' the chatbot responses were. Previous research has indicated similar results with a custom chatbot in response to COVID-19 [41]. However, as other studies did not measure overall user experience, there was insufficient evidence in the current review to generalise this benefit to other custom chatbots or infer positive user experience of non-custom chatbots such as ChatGPT when seeking AOD health information.

### 4.3 | Accuracy

Accuracy was measured in only one study [28] and garnered nuanced evaluation. The authors note that although the level of accuracy was promising, there were limitations, such as the lack of breadth and depth of response. In a separate study exploring the accuracy of chatbot responses when seeking medical information, it was found that ChatGPT-3.5 was acceptably accurate, but that accuracy tended to diminish as question difficulty increased [42], which has also been found when using ChatGPT to access dementia [43] and pharmaceutical drug [37] information. This has implications for public chatbot users as people should be able to access accurate, reliable information regardless of query complexity. Similarly, recent research has

shown that the accuracy of chatbots within oncology settings was variable, with information not aligned with current medical guidelines [44], which was echoed in one of the papers included in the current review [28]. Cabrera et al. [45] suggest that accuracy can be impacted by the quality of the prompt used and bias in both the user and LLM. As technology progresses, the accuracy and reliability of LLMs may improve with advancements such as using retrieval-augmented generation [46]. However, the mixed evidence on chatbot accuracy makes it difficult to conclude whether currently evaluated AI chatbots provide accurate AOD information for users.

#### 4.4 | Other Benefits

In the study exploring adequacy and appropriateness [24], both items were rated highly by the clinicians assessing ChatGPT-4 and LLaMA-2 outputs. However, in this study, the two domains were measured by only one item each with potential construct overlap between whether the chatbot response is adequate and whether it is appropriate. Therefore, these two benefits may reflect a similar outcome. Similarly, as a positive outcome, Russell et al. [27] found that ChatGPT mostly responded to alcohol use disorder questions with supporting documentation. The metric used was the number of peer-reviewed citations, which has construct overlap with evidence-based responses, which gained mixed results. Again, this highlights potential measurement and construct definition inconsistencies across studies.

#### 4.5 | Mixed Results

##### 4.5.1 | Evidence-Based Responses

Each item was measured in only one study, except for chatbot response quality and the extent to which responses were evidence based, each measured in two studies. In the two studies exploring whether chatbot responses were evidence based, the results were contradictory. Russell et al. [27] found that over 90% of responses to alcohol use disorder questions were fully evidence based, with the remaining responses partially evidence based. In support, medical research has shown that ChatGPT-4 is evidence based most of the time (90%) when asked urology questions [47], and 72% when asked neurosurgery questions [48]. In both medical-related studies, ChatGPT-4 outperformed ChatGPT3.5, and noteworthy, ChatGPT-4 performed better than most of the neurosurgeon participants [48]. Cardiology researchers in Germany have however warned against synonymising evidence-based chatbot responses with human expertise, as chatbots cannot compensate for a lack of first-hand clinical experience [49]. In the current review, Russell et al. [27] also advise against accepting LLM responses as medical advice. Presenting contrasting results from the included studies, Spallek et al. [28] found that although some evidence-based sources are used in ChatGPT-4's response to AOD questions, traditional, reputable health resources far outperform the chatbot on accuracy. The researchers concluded that ChatGPT used an insufficient number of scientific journal sources, and relied too heavily on lower quality sources [28]. With this review demonstrating

mixed results on whether chatbot responses on AOD topics are evidence based, further research is warranted.

#### 4.6 | Challenges in Chatbot Use for AOD Information-Seeking

##### 4.6.1 | Quality

Although 'quality' was not defined in either study [24, 28], one study reported the measurement item as 'What is the level of quality of the response?' (1 very poor, 2 below average, 3 average, 4 above average, 5 excellent) [24]. In the study that included a comparison of chatbots [24], LLaMA-2 outperformed ChatGPT-4, indicating potential variance in quality depending on the chatbot used. It should also be noted that the authors found that the clinician participants rated chatbot quality as good, despite potentially dangerous responses. This exemplifies the discrepancies between perceived and actual quality of chatbots, which can lead to people trusting chatbot information even when they should not. The authors also highlight the legal and health consequences of inaccurate information. The variability in quality and performance suggests that users need to be made aware of reliability issues.

##### 4.6.2 | Readability

In Spallek et al. [28], the chatbot responses were of varying quality, limited by characteristics such as readability, which has implications for people with low health literacy [8]. Other research has found that ChatGPT can make drug information more readable when prompted [50], however, in the current review, despite prompting the chatbot to produce a grade 8 response, the reading level remained higher than considered acceptable for general users [28]. Similarly, Ehlers et al. [44] found that cancer chatbot information is high quality but does not have broad readability, which is further supported in other literature [51, 52]. Requiring a high reading level to decipher the chatbot response may alienate users, making it inaccessible for those who do not have the skills and prior knowledge to evaluate and use the complex information provided by the chatbots.

##### 4.6.3 | User Safety

The safety of chatbot use was also a noted concern [24]. For example, ChatGPT inaccurately and against guidelines stated that people can safely and abruptly quit long-term heroin use by detoxing at home on 23% of occasions when asked [24]. For the public, low-quality chatbot responses may cause harm if inaccurate information is acted upon. A concern about the safety of chatbot information has been raised in the literature previously. In a study exploring physical activity and nutrition, young people reported concerns about being able to trust whether the chatbot information is false, misleading, exploitative or harmful [53]. This evidence contrasts with research on therapeutic interventions that have found chatbots are safe for substance use [54]. It has also been suggested that in a clinical setting, the speed and sensitivity of chatbot responses may

eventually lead to improved patient safety and reduced clinical errors [33]. However, with safety being such an important element of chatbot use, further evidence is needed to establish in what circumstances chatbot use is safe for AOD health information-seeking.

#### 4.6.4 | Stigmatising Language

People who use AOD are at risk of stigmatisation and may subsequently avoid seeking help [55]. Two of the studies included in the current review indicated stigmatising language within chatbot responses [24, 28]. This may have implications for those who are not yet ready to seek help in person due to stigma but may still want unbiased access to health information, without experiencing further shame. Previous research suggests that chatbots tend to be most acceptable for health conditions with lower levels of stigma [56], and therefore may be less applicable for AOD disorders, which tend to have high associated stigma [57]. In the current review, both studies that highlighted stigmatising language tested non-custom chatbots such as ChatGPT. It may be that custom health chatbots are less likely to include stigmatising language due to the way they are developed; however, further research is needed to explore this. Regardless of whether stigma is a factor in chatbot use, chatbots must provide AOD information for users in language that does not contribute further to the stigma of AOD use.

#### 4.7 | Gaps in the Literature

While AI is a burgeoning field of research generally, including in healthcare settings [58], the current AOD topic is much less investigated. Given the known harms associated with AOD use, more research is required to explore the potential of AI to reduce harms. Most studies used quantitative methods, highlighting a qualitative gap in the literature. This is an important gap as research on chatbots in healthcare has previously shown that despite favourable quantitative data about chatbot usability, qualitative data presented more mixed findings, with limitations such as readability, repetition and lack of interactivity not elucidated in the quantitative data [59]. This highlights the limitations of relying solely on quantitative data, especially with complex human phenomena such as risky AOD use.

Three papers did not include participants, indicating another gap in the evidence. Including participant perspectives on chatbot use would help us to understand the benefits and challenges that the public experiences when using chatbots, with the potential to investigate different cohorts, such as older versus younger people or those with low health literacy. With only five papers included, conducting future studies with various research designs is imperative. For example, no study designs in this review used experimental methods such as randomised control trials or longitudinal designs. A randomised control trial could investigate differences in the feasibility, acceptability and efficacy of various chatbots for AOD health information. In addition, further mixed methods and qualitative studies may provide richer and more balanced findings to help fill these gaps. Furthermore,

the included studies did not compare the performance of custom versus non-custom chatbots in seeking AOD health information. Studies that compare custom versus non-custom chatbot performance would be a beneficial contribution to the literature.

#### 4.8 | Strengths and Limitations

To our knowledge, this is the first review to explore AOD information-seeking with the public, offering a novel contribution to the literature. As with all literature reviews, the study is limited by the search terms used, which were comprehensive but not exhaustive, and by the databases available to the team. As the study included English language papers only, the study is potentially missing evidence from non-English speaking authors. The study aimed to map empirical data and therefore excluded grey literature; however, this exclusion criteria may have narrowed the results further. The application of the inclusion and exclusion criteria resulted in a final set of only five papers for analysis. While scoping reviews do not require a minimum number of studies, this small set still provided valuable insights. The selected papers offered a diverse range of AI-supported chatbots addressing AOD-related health information, allowing key themes to be identified and explored.

#### 4.9 | Implications and Future Research

While research in AI is experiencing rapid growth, the results from this scoping review indicate a substantial gap in the literature. There are currently insufficient studies examining the use of LLMs to obtain AOD health information. Further studies are needed to understand the benefits, risks, ethical implications. Though there was some construct overlap, the included articles were heterogeneous in the items measured, limiting the comparability across studies. Future research could explore which evaluation domains are commonly used and which are most useful in assessing chatbot response quality and make recommendations for core evaluation domains. Future studies could incorporate this set of consistent measures, and employ the use of multiple raters for reliability purposes, so that there is improved comparability across studies.

The topic of accuracy and safety also warrants deeper investigation from two distinct perspectives: custom vs. non-custom chatbots; and the different meanings of safety dependent on the context of the query provided. Given that ChatGPT and other non-custom chatbots are not within the control of health organisations, this might influence recommendations for public chatbot use, for example, encouraging people to use custom chatbots that have been evaluated as safe and accurate for health information. Finally, as both custom and non-custom chatbots are entering the market quickly, there will be a notable lag in the evidence base that should be considered when providing recommendations on chatbot use.

Driven by the rapidly evolving nature of LLM chatbots and building on the future research directions outlined above, further investigation is warranted into how these tools might be extended to support treatment approaches for AOD use. Realising this area of growing significance will most likely require addressing



the same critical issues of accuracy, safety, readability, evidence base and quality—representing a vital new body of work.

## 5 | Conclusion

This scoping review summarised the literature regarding LLM chatbots for AOD information-seeking. It is evident that chatbot use is a growing field of research and consequently has some gaps in evidence, particularly qualitative studies and authorship by a broader researcher origin. There is a need to use consistent measures when evaluating chatbot use to ensure that studies can be compared and synthesised. The current review suggests that gaps in knowledge remain in the areas of accuracy, safety, readability, evidence base and quality of AI chatbot responses to AOD questions. Given the known harms associated with AOD use, the need to evaluate the safety of chatbot responses to AOD information seeking is paramount. Future research should aim to include participants, therefore addressing the identified gaps in qualitative, mixed methods and longitudinal studies. Research should investigate whether LLMs can provide accurate AOD information that adheres to non-stigmatising, person-centred language and specific reading levels to ensure accessibility.

### Author Contributions

**Natasha Harding:** conceptualisation, methodology, analysis, writing – original draft, writing – review and editing. **Nataly Bovopoulos:** conceptualisation, methodology, analysis, writing – original draft, writing – review and editing. **Dotahn Caspi:** conceptualisation, writing – original draft, writing – review and editing. **Craig Martin:** writing – review and editing. **Skye McPhie:** writing – review and editing. **Mohamed Reda Bouadjenek:** supervision, writing – review and editing. **Sunil Aryal:** supervision, writing – review and editing. **Michael Hobbs:** supervision, writing – review and editing. Each author certifies that their contribution to this work meets the standards of the International Committee of Medical Journal Editors.

### Acknowledgements

The authors thank Linda Gay, ADF Specialist Librarian and Information Officer, for assistance developing the search strings and accessing ADF library resources. We sincerely thank the reviewers for their insightful comments and constructive feedback, which greatly contributed to improving the quality and clarity of this manuscript.

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

### References

1. L. Degenhardt, F. Charlson, A. Ferrari, et al., “The Global Burden of Disease Attributable to Alcohol and Drug Use in 195 Countries and Territories, 1990–2016: A Systematic Analysis for the Global Burden of Disease Study 2016,” *Lancet Psychiatry* 5, no. 12 (2018): 987–1012.
2. Australian Institute of Health and Welfare, National Drug Strategy Household Survey 2022–2023, 2024, <https://www.aihw.gov.au/reports/illegal-use-of-drugs/national-drug-strategy-household-survey>.

3. Australian Institute of Health and Welfare, Alcohol, Tobacco & Other Drugs in Australia, 2023, <https://www.aihw.gov.au/reports/alcohol/alcohol-tobacco-other-drugs-australia/contents/impacts>.
4. J. Kavitha, S. Sivakrishnan, and N. Srinivasan, “Self Medication in Today’s Generation Without Knowledge as Self Inflicted Harm,” *Archives of Pharmacy Practice* 13, no. 3 (2022): 16–22.
5. Alcohol and Drug Foundation, *Alcohol and Other Drugs Information. Survey Report* (Alcohol and Drug Foundation, 2022).
6. L. Sbaffi and J. Rowley, “Trust and Credibility in Web-Based Health Information: A Review and Agenda for Future Research,” *Journal of Medical Internet Research* 19, no. 6 (2017): e218.
7. A. M. Mason, J. Compton, and S. Bhati, “Disabilities and the Digital Divide: Assessing Web Accessibility, Readability, and Mobility of Popular Health Websites,” *Journal of Health Communication* 26, no. 10 (2021): 667–674.
8. N. Diviani, B. van den Putte, S. Giani, and J. C. van Weert, “Low Health Literacy and Evaluation of Online Health Information: A Systematic Review of the Literature,” *Journal of Medical Internet Research* 17, no. 5 (2015): e112.
9. M. Al-Amin, M. S. Ali, A. Salam, et al., “History of Generative Artificial Intelligence (AI) Chatbots: Past, Present, and Future Development,” *arXiv* (2024): 240205122.
10. E. Jo, D. A. Epstein, H. Jung, and Y.-H. Kim, Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. CHI Conference on Human Factors in Computing Systems, 2023. p. 1–16.
11. A. G. Dunn, I. Shih, J. Ayre, and H. Spallek, “What Generative AI Means for Trust in Health Communications,” *Journal of Communication in Healthcare* 16, no. 4 (2023): 385–388.
12. M. Sallam, “ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns,” *Healthcare* 11, no. 6 (2023): 887, <https://doi.org/10.3390/healthcare11060887>.
13. F. Muftić, M. Kadunić, A. Mušibegović, and A. Abd Almisreb, “Exploring Medical Breakthroughs: A Systematic Review of ChatGPT Applications in Healthcare,” *Southeast Europe Journal of Soft Computing* 12, no. 1 (2023): 13–41.
14. J. L. Freeman, P. H. Y. Caldwell, P. A. Bennett, and K. M. Scott, “How Adolescents Search for and Appraise Online Health Information: A Systematic Review,” *Journal of Pediatrics* 195 (2018): 244–55.e1.
15. L. De Angelis, F. Baglivo, G. Arzilli, et al., “ChatGPT and the Rise of Large Language Models: The New AI-Driven Infodemic Threat in Public Health,” *Frontiers in Public Health* 11 (2023): 1166120.
16. T. Templin, M. W. Perez, S. Sylvia, J. Leek, and N. Sinnott-Armstrong, “Addressing 6 Challenges in Generative AI for Digital Health: A Scoping Review,” *PLoS Digital Health* 3, no. 5 (2024): e0000503.
17. S. Harrer, “Attention Is Not All You Need: The Complicated Case of Ethically Using Large Language Models in Healthcare and Medicine,” *eBioMedicine* 90 (2023): 104512.
18. A. M. Afsahi, S. A. S. Alinaghi, A. Molla, et al., “Chatbots Utility in Healthcare Industry: An Umbrella Review,” *Frontiers in Health Informatics* 13 (2024): 200.
19. L. Ogilvie, J. Prescott, and J. Carson, “The Use of Chatbots as Supportive Agents for People Seeking Help With Substance Use Disorder: A Systematic Review,” *European Addiction Research* 28, no. 6 (2022): 405–418.
20. N. S. Bonfiglio, M. L. Mascia, S. Cataudella, and M. P. Penna, “Digital Help for Substance Users (SU): A Systematic Review,” *International Journal of Environmental Research and Public Health* 19, no. 18 (2022): 11309.
21. E. M. Boucher, N. R. Harake, H. E. Ward, et al., “Artificially Intelligent Chatbots in Digital Mental Health Interventions: A Review,” *Expert Review of Medical Devices* 18 (2021): 37–49.

22. A. Aggarwal, C. C. Tam, D. Wu, X. Li, and S. Qiao, "Artificial Intelligence-Based Chatbots for Promoting Health Behavioral Changes: Systematic Review," *Journal of Medical Internet Research* 25 (2023): e40789.
23. M. Peters, C. Godfrey, P. McInerney, Z. Munn, A. Tricco, and H. Khalil, "Chapter 11: Scoping Reviews," in *JBIR Reviewer's Manual*, ed. E. Aromataris and Z. Munn (JBI, 2020).
24. S. Giorgi, K. Isman, T. Liu, Z. Fried, J. Sedoc, and B. Curtis, "Evaluating Generative AI Responses to Real-World Drug-Related Questions," *Psychiatry Research* 339 (2024): 116058.
25. K. Loveys, E. Lloyd, M. Sagar, and E. Broadbent, "Development of a Virtual Human for Supporting Tobacco Cessation During the COVID-19 Pandemic," *Journal of Medical Internet Research* 25 (2023): e42310.
26. M. G. Monteiro, D. Pantani, I. Pinsky, and T. A. H. Rocha, "Using the Pan American Health Organization Digital Conversational Agent to Educate the Public on Alcohol Use and Health: Preliminary Analysis," *JMIR Formative Research* 7, no. 1 (2023): e43165.
27. A. M. Russell, S. F. Acuff, J. F. Kelly, J.-P. Allem, and B. G. Bergman, "ChatGPT-4: Alcohol Use Disorder Responses," *Addiction* 119, no. 12 (2024): 2205–2210.
28. S. Spallek, L. Birrell, S. Kershaw, E. K. Devine, and L. Thornton, "Can We Use ChatGPT for Mental Health and Substance Use Education? Examining Its Quality and Potential Harms," *JMIR Medical Education* 9, no. 1 (2023): e51243.
29. H. Baber, K. Nair, R. Gupta, and K. Gurjar, "The Beginning of ChatGPT: A Systematic and Bibliometric Review of the Literature," *Information and Learning Science* 125, no. 7/8 (2024): 587–614.
30. Z. Deng, W. Ma, Q.-L. Han, et al., "Exploring DeepSeek: A Survey on Advances, Applications, Challenges and Future Directions," *IEEE/CAA Journal of Automatica Sinica* 12, no. 5 (2025): 872–893.
31. A. Saabith, T. Vinothraj, and M. Fareez, "Exploring the Landscape of Large Language Models: A Comprehensive Review of Current Technologies," *International Journal of Research in Engineering and Science* 12, no. 12 (2024): 166–188.
32. K. Abu Hammour, H. Alhamad, F. Y. Al-Ashwal, A. Halboup, R. Abu Farha, and A. Abu Hammour, "ChatGPT in Pharmacy Practice: A Cross-Sectional Exploration of Jordanian Pharmacists' Perception, Practice, and Concerns," *Journal of Pharmaceutical Policy and Practice* 16, no. 1 (2023): 115.
33. F. Y. Al-Ashwal, M. Zawiah, L. Gharaibeh, R. Abu-Farha, and A. N. Bitar, "Evaluating the Sensitivity, Specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard Against Conventional Drug-Drug Interactions Clinical Tools," *Drug, Healthcare and Patient Safety* 15 (2023): 137–147.
34. T. Daniel, A. de Chevigny, A. Champrigaud, et al., "Answering Hospital Caregivers' Questions at Any Time: Proof-Of-Concept Study of an Artificial Intelligence-Based Chatbot in a French Hospital," *JMIR Human Factors* 9, no. 4 (2022): e39102.
35. J. Koman, K. Fauvelle, S. Schuck, N. Texier, and A. Mebarki, "Physicians' Perceptions of the Use of a Chatbot for Information Seeking: Qualitative Study," *Journal of Medical Internet Research* 22, no. 11 (2020): e15185.
36. F. Montastruc, W. Storck, C. de Canecaude, et al., "Will Artificial Intelligence Chatbots Replace Clinical Pharmacologists? An Exploratory Study in Clinical Practice," *European Journal of Clinical Pharmacology* 79, no. 10 (2023): 1375–1384.
37. B. Morath, U. Chiriac, E. Jaskowski, et al., "Performance and Risks of ChatGPT Used in Drug Information: An Exploratory Real-World Analysis," *European Journal of Hospital Pharmacy* 31, no. 6 (2023): 491–497.
38. R. Mosleh, Q. Jarrar, Y. Jarrar, M. Tazkarji, and M. Hawash, "Medicine and Pharmacy Students' Knowledge, Attitudes, and Practice Regarding Artificial Intelligence Programs: Jordan and West Bank of Palestine," *Advances in Medical Education and Practice* 14 (2023): 1391–1400.
39. M. S. Sheikh, E. F. Barreto, J. Miao, et al., "Evaluating Chatgpt's Efficacy in Assessing the Safety of Non-Prescription Medications and Supplements in Patients With Kidney Disease," *Digital Health* 10 (2024): 10.
40. M. Laymouna, Y. Ma, D. Lessard, T. Schuster, K. Engler, and B. Leblouché, "Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review," *Journal of Medical Internet Research* 26 (2024): e56930.
41. B. A. Chagas, A. S. Pagano, R. O. Prates, et al., "Evaluating User Experience With a Chatbot Designed as a Public Health Response to the COVID-19 Pandemic in Brazil: Mixed Methods Study," *JMIR Human Factors* 10 (2023): e43135.
42. R. S. Goodman, J. R. Patrinely, J. C. A. Stone, et al., "Accuracy and Reliability of Chatbot Responses to Physician Questions," *JAMA Network Open* 6, no. 10 (2023): e2336483.
43. H. R. Saeidnia, M. Kozak, B. D. Lund, and M. Hassanzadeh, "Evaluation of Chatgpt's Responses to Information Needs and Information Seeking of Dementia Patients," *Scientific Reports* 14, no. 1 (2024): 10273.
44. J. Ehlers, "AI Chatbots Can Provide Accurate Cancer Information but Have Limitations," *Monthly Prescribing Reference* (2023).
45. J. Cabrera, M. S. Loyola, I. Magaña, et al., "Ethical Dilemmas, Mental Health, Artificial Intelligence, and LLM-Based Chatbots," in *Lecture Notes in Computer Science*, vol. 13920 (Springer, 2023), 313–326.
46. R. Yang, Y. Ning, E. Keppo, et al., "Retrieval-Augmented Generation for Generative Artificial Intelligence in Health Care," *npj Health Systems* 2, no. 1 (2025): 2.
47. Z. Zhou, X. Wang, X. Li, and L. Liao, "Is ChatGPT an Evidence-Based Doctor?," *European Urology* 84, no. 3 (2023): 355–356.
48. J. Liu, J. Zheng, X. Cai, D. Wu, and C. Yin, "A Descriptive Study Based on the Comparison of ChatGPT and Evidence-Based Neurosurgeons," *iScience* 26, no. 9 (2023): 107590.
49. W. Haverkamp, J. Tennenbaum, and N. Strodthoff, "ChatGPT Fails the Test of Evidence-Based Medicine," *European Heart Journal - Digital Health* 4, no. 5 (2023): 366–367.
50. A. Juhi, N. Pipil, S. Santra, S. Mondal, J. K. Behera, and H. Mondal, "The Capability of ChatGPT in Predicting and Explaining Common Drug-Drug Interactions," *Cureus* 15, no. 3 (2023): e36272.
51. I. Ulusoy, M. Yilmaz, and A. Kivrak, "How Efficient Is ChatGPT in Accessing Accurate and Quality Health-Related Information?," *Cureus* 15, no. 10 (2023): e46662.
52. W. Andrikyan, S. M. Sametinger, F. Kosfeld, et al., "Artificial Intelligence-Powered Chatbots in Search Engines: A Cross-Sectional Study on the Quality and Risks of Drug Information for Patients," *BMJ Quality and Safety* 34 (2024): 017476.
53. R. Han, A. Todd, S. Wardak, S. R. Partridge, and R. Raeside, "Feasibility and Acceptability of Chatbots for Nutrition and Physical Activity Health Promotion Among Adolescents: Systematic Scoping Review With Adolescent Consultation," *JMIR Human Factors* 10 (2023): e43227.
54. J. J. Prochaska, E. A. Vogel, A. Chieng, et al., "A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study," *Journal of Medical Internet Research* 23, no. 3 (2021): e24850.
55. S. W. Finn, A. Mejdal, and A. S. Nielsen, "Public Stigma and Treatment Preferences for Alcohol Use Disorders," *BMC Health Services Research* 23, no. 1 (2023): 76.
56. O. Miles, R. West, and T. Nadarzynski, "Health Chatbots Acceptability Moderated by Perceived Stigma and Severity: A Cross-Sectional Survey," *Digital Health* 7 (2021): 20552076211063012.

57. S. M. Rundle, J. A. Cunningham, and C. S. Hendershot, "Implications of Addiction Diagnosis and Addiction Beliefs for Public Stigma: A Cross-National Experimental Study," *Drug and Alcohol Review* 40, no. 5 (2021): 842–846.

58. R. K. Garg, V. L. Urs, A. A. Agarwal, S. K. Chaudhary, V. Paliwal, and S. K. Kar, "Exploring the Role of ChatGPT in Patient Care (Diagnosis and Treatment) and Medical Research: A Systematic Review," *Health Promotion Perspective* 13, no. 3 (2023): 183–191.

59. M. Milne-Ives, C. de Cock, E. Lim, et al., "The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review," *Journal of Medical Internet Research* 22, no. 10 (2020): e20346.

Appendix A

Search strings for each database

ADF Library

('large language model' OR LLM OR AI OR 'artificial intelligence' OR GPT OR 'natural language processing' OR NLP AND chat\*) AND (drug OR alcohol OR substance OR dependen\* OR addict\* OR smok\* OR vap\* OR AOD) AND (health AND information NOT detection NOT 'health record\*' NOT surveillance)

Google Scholar

information AND alcohol OR drug OR substance OR AOD OR addiction OR dependence AND chatbot

PsychArticles

((('large language model' or LLM or AI or 'artificial intelligence' or GPT or 'natural language processing' or NLP or 'machine learning') and chat\* and (drug or alcohol or substance or dependen\* or addict\* or smok\* or vap\* or AOD) and (health and information)) not detection not 'health record\*' not surveillance).mp. [mp = title, abstract, full text, caption text]

PubMed

('large language model'[Title] OR LLM[Title] OR chatbot[Title]) AND information AND (alcohol OR drug OR AOD)

Appendix B

See Table A1

TABLE A1 | Blank data extraction form.

Autor/year
Country of origin
Aim/purpose
Type of study
Population/participants
AI/chatbot
AOD
Information seeking
Key findings relating to this scoping review

Abbreviation: AOD, alcohol and other drugs.