MUNI FSS

JOURNAL OF PSYCHOSOCIAL RESEARCH ON CYBERSPACE
CYBERPSYCHOLOGY

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2026). AI aversion? Effects of author disclosure on young people's perceptions of mental health advice. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 20*(2), Article 1. https://doi.org/10.5817/CP2026-2-1

AI Aversion? Effects of Author Disclosure on Young People's Perceptions of Mental Health Advice

Petter Bae Brandtzaeg[1,2], *Marita Skjuve*[1], *& Asbjørn Følstad*[1]

[1] SINTEF Digital, Oslo, Norway
[2] Department of Media and Communication, University of Oslo, Oslo, Norway

Abstract

The increasing use of large language models (LLMs), such as ChatGPT, is already impacting how young people seek mental health support online. However, AI aversion, the reluctance or resistance individuals feel toward AI, may influence individuals' perceptions and willingness to engage with LLM-generated advice. In this mixed-method study, we investigated how 440 young people (aged 17–21) perceived mental health advice from ChatGPT compared with that of health professionals, emphasizing the effect of author disclosure. Participants assessed answers from ChatGPT and health professionals across four dimensions—Validation, Relevance, Clarity, and Utility—and were asked to recommend answers. The findings indicate a preference for AI-generated answers when participants were unaware of the author's identity: ChatGPT's answers scored significantly higher on Validation, Relevance, Clarity, and Utility. Conversely, when the author was disclosed, participants favored responses from health professionals and rated their answers significantly higher for Validation, indicating AI aversion. Qualitative data further revealed that participants became more critical when they knew the content was AI-generated, while responses from health professionals were viewed as more credible, empathetic, and tailored. These findings may indicate human favoritism. The study makes the key contribution of identifying how source awareness impacts the reception of AI-generated content in a sensitive domain. To address the potential for AI aversion within help-seeking, our findings suggest the importance of developing hybrid human–AI support models that combine the efficiency of AI with the relational legitimacy of human professionals, improving both the acceptance and impact of digital mental health support.

Keywords: ChatGPT; LLM; AI aversion; youth; mental health support

Editorial Record

First submission received:
May 6, 2025

Revisions received:
January 30, 2026
March 10, 2026

Accepted for publication:
March 12, 2026

Editor in charge:
Emmelyn Croes

Introduction

AI aversion concerns the tendency to distrust or reject outputs generated by artificial intelligence (AI), even when an AI performs as well as or better than human alternatives (Dietvorst et al., 2015). This tendency represents a growing challenge for the adoption of AI-driven technologies in sensitive domains such as mental health support (Arvai et al., 2025). This skepticism is often reinforced by the opaque, "black box" nature of large language models (LLMs), which limits users' ability to assess the reliability and trustworthiness of AI-generated advice (Joyce et al.,

2023). At the same time, general purpose LLMs such as ChatGPT, Claude, DeepSeek, and Gemini offer significant promise in digital help-seeking contexts (Khurana et al., 2024; Mendel et al., 2025). These tools can produce interactive, personalized, and immediate responses (Brown et al., 2020), and may help bridge accessibility gaps for young people who face barriers to seeking support, such as fear of stigma (Pretorius et al., 2019), financial constraints, and distrust of formal institutions (Hudson et al., 2010). Insofar as AI aversion leads to reduced use of general purpose LLMs, this could strengthen unfortunate divides in skill levels needed to benefit from the technology (Kacperski et al., 2025).

AI aversion may indeed serve as a protective heuristic in high-stakes domains (De-Arteaga et al., 2020) such as mental health. At the same time, AI aversion could hamper the potential benefits of scalable LLM-based help-seeking. Importantly, the value of LLMs as mental health support tools depends on the quality of the advice they can provide. Particularly, whether they can produce advice perceived as validating, relevant, applicable, and easy to understand. These attributes are important in the context of mental healthcare across all modes, whether computer-mediated (Bickmore et al., 2005) or face-to-face (Duncan et al., 2003). In addition, users' perceptions of source credibility play a role in determining whether health advice is trusted and acted upon (Gaskin et al., 2024; Khurana et al., 2024).

While prior research has explored the impact of AI disclosure and AI aversion in other contexts, such as relationship counseling (Vowels, 2024) and advice-seeking in shopping (Dang & Liu, 2024), little work has examined AI aversion in the context of AI-generated mental health advice and help-seeking. In particular, there is little empirical evidence comparing how young people perceive and evaluate mental health information generated by LLMs when the information's AI provenance is disclosed versus undisclosed. Addressing this research gap is critical, as more young people seek mental health support not only from online services increasingly based, at least in part, on LLMs (e.g., Woebot, Wysa, Koko, Youper, Replika) but also from general purpose LLMs such as ChatGPT and Gemini (Brandtzæg et al., 2021).

Therefore, we investigate how young people experience and evaluate answers generated from LLMs compared to those authored by human professionals and whether awareness of the source (AI vs. human) influences their perceptions. Understanding these perceptions is key to assessing whether and how young people prefer support on mental health from humans or AI. We approach this by explicating the following research questions:

RQ1: How do young people perceive and evaluate mental health advice in help-seeking contexts when comparing answers provided by an LLM vs. human health professionals?

RQ2: Which attributes of mental health advice are perceived by young people as key when evaluating answers provided by an LLM vs. human health professionals?

RQ3: How does source awareness (AI vs. human) influence young people's mental health advice preferences?

To address these research questions, we conducted a mixed-methods investigation using open-ended and closed items in a questionnaire-based experiment, in which 440 young Norwegians (aged 17–21) were randomly assigned to one of two conditions. In the first condition, the participants were informed about the source of the mental health advice they received (AI vs. a health professional), in the second condition, the participants did not. In both conditions, the participants were asked to evaluate answers to two questions on mental health issues taken from an online help service offering anonymous advice to young people. Each of the two questions had two associated answers – one generated by an LLM and one authored by a health professional. Hence, the participants evaluated four answers in total. The participants were asked to rate each answer in terms of four dimensions adapted from the *Session Rating Scale* (Duncan et al., 2003), a validated tool for assessing perceived quality in therapeutic interactions: perceived Validation, Relevance, Clarity, and Utility. Furthermore, the participants were asked to elaborate on their evaluation in an open-ended response, allowing us to identify the attributes of the answer perceived as key.

By incorporating user-generated questions from an online help service and involving potential young users as participants, this study offers valuable insights into how perceptions of authorship shape human–AI interaction in sensitive contexts. These findings specifically inform the design of digital mental health support. Understanding young people's perceptions of AI-generated advice is crucial for evaluating the conditions under which LLM-based help services can achieve successful adoption and sustained engagement. Additionally, understanding how transparency about AI's role influences willingness to engage with mental health advice can help policymakers and developers establish ethical guidelines for AI use in sensitive contexts such as mental health care. While it will typically be important to disclose the source of mental health advise, for ethical as well as regulatory reasons,

understanding the potential implications of doing so is critical – in particular to understand any impact of AI aversion. More broadly, these results are relevant to any domain where source credibility is critical to user engagement with AI.

Background

There is a clear tendency for people, especially young people, to increasingly use LLMs for social support in general and in the context of mental health (Brandtzæg et al., 2021; Brandtzaeg et al., 2025). Moreover, help services are experimenting with using LLMs to support young people by delivering efficient 24/7 responses (Khurana et al., 2024). This trend is driven by LLMs like ChatGPT, which are highly sophisticated and capable of empathetic interactions on virtually any topic, in any language, at any time (Brown et al., 2020; Sorin et al., 2024).

Several studies have explored LLMs for mental health support (Hua et al., 2024; Jiang et al., 2024; T. Kim et al., 2024; Na, 2024), where these show promise as tools for supporting people with mental health-related concerns. Research has also examined public perceptions of AI-generated advice versus advice from human experts, especially in healthcare contexts (Ayers et al., 2023; Durairaj et al., 2024; Singhal et al., 2025; Small et al., 2024). Here, research tends to show how AI advice is often rated more favorably than human-written (Ayers et al., 2023; J. Kim et al., 2024).

According to Ayers et al. (2023), healthcare practitioners rated AI-generated medical advice more favorably than that of physicians, noting that the AI-produced responses were of higher quality and demonstrated greater empathy. In a related study, Singhal et al. (2025) observed that while AI-generated responses to medical inquiries were favored over those from general practitioners, they did not surpass the quality of answers provided by specialists. Vowels (2024) investigated ChatGPT's capacity to offer relationship counselling and discovered that participants viewed its responses as more empathetic and useful than those from relationship experts. Similarly, Hatch et al. (2025) found that ChatGPT could produce therapeutic responses that were frequently indistinguishable from those crafted by licensed therapists, and the AI-generated responses scored higher on adherence to core psychotherapy principles.

Thus far, most research comparing AI-generated advice to human experts has focused on older users, with the exception of a study by Young et al. (2024), who examined young people. This study showed that participants strongly preferred advice from adult mentors overall, ranking AI-generated responses second. However, in that study, preferences for AI varied by topic: for highly sensitive issues such as suicidal ideation, participants favored adult mentors, while AI was preferred for less emotionally charged topics.

Despite recent research, studies focusing on young people's mental health remain scarce. Because perceptions of advice differ depending on whether it is AI- or human-generated (Vowels, 2024), understanding how young people evaluate LLM-based support is increasingly important.

AI Aversion

AI aversion is central to human–AI interaction in help-seeking contexts because reliance on LLMs depends not only on the quality of advice but also on trust in AI-generated content. Thus, even if LLM answers are perceived as empathetic or informative, as several studies suggest, user perceptions of those answers may change significantly once they know the content is AI-generated. Such effects are therefore important to understand as help services explore the transformational potential of LLMs.

The notion of AI aversion derives from the concept of *algorithmic aversion* introduced by Dietvorst et al. (2015). Algorithmic aversion describes a phenomenon whereby individuals distrust or are reluctant to rely on algorithmic or automated recommendations, predictions, or decisions despite evidence that algorithms may be, at least in some instances, accurate, objective, or efficient (Dietvorst et al., 2015). This aversion may lead individuals to favor human input even in contexts where AI performs better (Castelo et al., 2019; Dietvorst et al., 2015).

Algorithmic aversion has been extensively studied across contexts ranging from health care to customer service (Castelo et al., 2019; Jussupow et al., 2020). In a literature review, Jussupow et al. (2020) identified four key algorithm-related characteristics that influence algorithmic aversion:

Algorithm agency: Whether the algorithm acts autonomously (performative) or provides advice (advisory).

Algorithm performance: The algorithm's reliability, accuracy, and error rates.

Perceived capabilities: Users' perceptions of the algorithm's ability to perform the task effectively.

Human involvement: The extent to which humans are involved in the algorithm's development or operation.

The term AI aversion has gained traction more recently, interpreted as a specification of algorithmic aversion reflecting the growing use and influence of generative AI systems. Recent research has documented AI aversion to LLM outputs. For instance, Vowels (2024) discovered that participants rated responses as less empathetic when they believed an LLM had generated the content. Böhm et al. (2023) showed that awareness of the author's identity (AI vs. human) influenced perceived competence, with LLM authors viewed as less competent when their identity was disclosed. They also found no significant differences in perceived usefulness or willingness to follow or share advice across sources or authorship.

Similarly, Proksch et al. (2024) found that participants perceived LLM authors as less competent than human authors when aware of the AI's identity. Additionally, Henestrosa and Kimmerle (2024) and Osborne and Bailey (2025) have demonstrated that knowledge that LLMs are the source of advice negatively influences perceptions of advice quality. Zhang and Gosline (2023) present a complementary perspective, showing that perceived quality increases when human involvement is known, while AI involvement has no effect, thus indicating *human favoritism* rather than AI aversion.

This research suggests that user perceptions of AI, particularly AI aversion or human favoritism, may impact AI-powered mental health support. Hence, the potential of LLMs to offer scalable, anonymous support that may lower barriers to entry for youth needs to be understood in the context of users' perceptions of AI. Understanding whether and how young people display AI aversion in sensitive contexts, such as mental health support, is crucial for integrating AI into support services.

Methods

Study Design and Context

To address our research questions, we employed a mixed-methods experimental design, comparing two groups of young participants under different conditions designed to assess how they evaluate mental health advice provided by ChatGPT (GPT-4) and health professionals. Participants in condition Source Known were informed about the source of the advice (AI vs. human), while participants in condition Source Unknown were not privy to this information. The dependent variables were the participants' evaluations of the advice-providing answers in terms of Validation, Relevance, Clarity, and Utility (Simpson et al., 2022). Participants were also asked to recommend answers and freely report on their motivations for making recommendations; this provided qualitative insight into their evaluations.

The study procedure is illustrated in Figure 1 and described in detail below. To ensure transparency, details about the study, including the methods, survey questionnaire, and ethical considerations, are available on the Open Science Framework (OSF; https://osf.io/x7cw9). The data collection was conducted in June 2024.

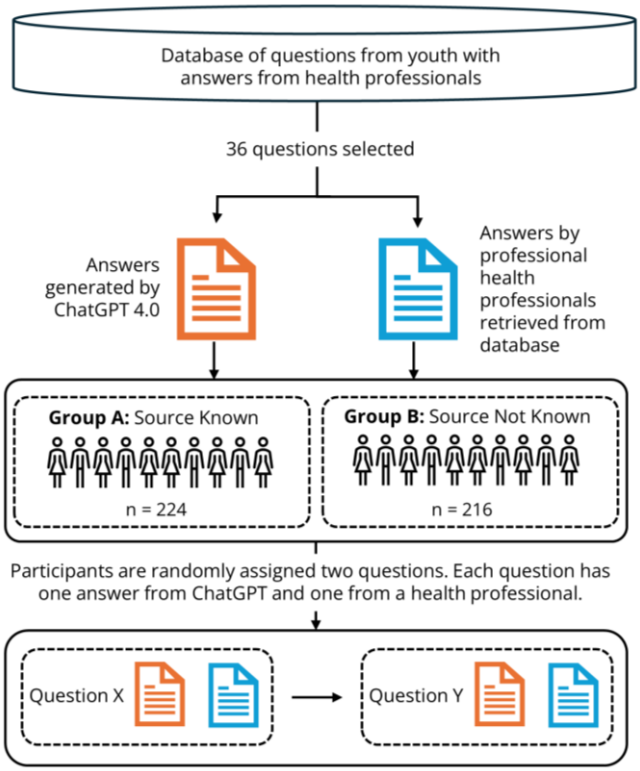
The study was conducted in collaboration with *ung.no*, an open, free, government-funded Norwegian help service that offers anonymous advice to young people aged 13–20. *ung.no* receives over 100,000 questions annually on topics such as mental health, family, relationships, sexuality, and health. All questions received by ung.no are answered by health professionals within 1–5 days. For transparency and public benefit, questions and answers are published openly on the website. We have drawn the questions and corresponding health professional answers used in this study from here (see Figure 1).

Participant Recruitment and Representativeness

A total of 667 young Norwegians aged 17–21 were recruited in June 2024 through Ipsos Survey Company[i], using an online, stratified sampling method to ensure demographic balance across gender and geographic location (urban vs. rural). Following quality checks, 226 participants (34%) were excluded due to inadequate responses to open-ended questions, typically blank or nonsensical replies. The final sample comprised 441 participants (51% female, 49% male; 35% rural, 65% urban), reflecting a balanced demographic distribution. In terms of AI usage, 43% of the participants reported using ChatGPT weekly or more, and 14% reported using Snapchat's My AI with

the same frequency, reflecting varying levels of exposure to different generative AI tools among young Norwegians.

Figure 1. *Overview of the Process for Gathering and Presenting Questions and Answers to Participants.*



Procedure

Selection of Questions

To ensure a representative and valid comparison, we incorporated real user-generated help-seeking questions into the study design. Specifically, we selected a diverse subset of 36 anonymized youth-generated questions from *ung.no*, drawn from a broader pool of 324 mental health-related inquiries (see Figure 1). This approach enhanced the study's relevance and realism. Example questions and the corresponding answers from both health professionals and ChatGPT are publicly available on the OSF.

We applied the following criteria when selecting the help-seeking questions: (1) questions should represent common and relevant mental health issues for young people seeking support, (2) questions should not contain topics that can be upsetting to the participants, such as questions regarding eating disorders, suicide or sexual abuse, (3) questions should not contain specific information that could identify the participants, such as city or residence or parents' profession, (4) questions should be submitted after September 2021 to avoid inclusion in the ChatGPT training data at the time of the study, and (5) questions should be gender balanced to ensure diverse representation.

Generation of Answers

Answers from health professionals to the 36 questions were retrieved from the *ung.no* database, that is, those provided online by *ung.no* in response to the posted questions. Answers from ChatGPT to the same 36 questions were generated in May 2023 using ChatGPT 4.0 via OpenAI's web interface (chat.openai.com). No additional instructions were included in the prompts beyond inputting each question. We cleared the prompt history before generating each answer to avoid contextual bias and cross-contamination between answers. Furthermore, chat history was disabled to prevent data retention by the system, an OpenAI user setting that allows users to opt out of having their data used to train or improve the model. All questions and answers from *ung.no*, as well as all ChatGPT-generated answers, were in Norwegian.

Participant Assignment to Conditions

Each participant was randomly assigned to one of two study conditions: Source Unknown or Source Known. This between-subjects design allowed us to isolate the effect of source disclosure on the evaluation of mental health advice and to examine the role of AI aversion in human–AI interaction and help-seeking contexts.

Condition: Source Unknown (n = 224). The participants in this condition did not know the identity of the source of the answers to either question. We only stated that the answers were from an "online help service." We removed identifiable information specific to health professionals at *ung.no* (e.g., "Thanks for reaching out to ung.no") and ChatGPT (e.g., "As an AI, I cannot provide a diagnosis"). This condition was established to examine participants' evaluations independent of source-based expectations or biases.

Condition: Source Known (n = 216). The participants in this condition were told the source identity of the answers to both questions. Answers were labelled "health professional" for the answers provided by humans and "ChatGPT (conversational robot/chatbot)" for the AI-generated answers. This condition allowed us to assess source disclosure and how knowledge that an answer was produced by an LLM influenced participants' experience.

Question Assignment and Answer Evaluation

After being assigned to a condition, the participants were randomly assigned two questions, each with a corresponding pair of answers: one authored by a health professional and one generated by ChatGPT. Hence, each participant was assigned to four answers in total. We administered the same questions and answers in both conditions to facilitate direct comparisons. For both conditions, data were collected using the LimeSurvey questionnaire tool, as outlined below.[1]

Participants evaluated each of the four answers assigned to them on four dimensions adapted from the *Session Rating Scale* (Duncan et al., 2003), a validated tool for assessing perceived quality in therapeutic interactions. The original Session Rating Scale was adapted to suit the context of AI-generated and human-authored mental health advice, focusing on dimensions relevant to online help-seeking. Each dimension was rated on a seven-point Likert scale, ranging from *strongly disagree* (1) to *strongly agree* (7). We chose to apply only one item per dimension to avoid survey fatigue.

For each participant, a composite score was calculated for each author type (i.e., health professionals or ChatGPT) by averaging their evaluations of the two answers provided by that author. This approach allowed us to capture a more stable and representative perception of each source, minimizing potential noise or bias from individual answer variability. The four evaluation dimensions and their corresponding questionnaire items are presented in Table 1.

Table 1. *Dimensions for Answer Evaluation and Corresponding Questionnaire Items.*

Dimension	Questionnaire item
Validation	The answer makes the person asking feel seen and heard
Relevance	The answer is relevant
Clarity	The answer is easy to understand
Utility	The answer is useful

Answer Recommendation

After evaluating two answers to a given help-seeking question, participants were asked to review both answers and indicate whether they would recommend either or both. This step was included to assess participants' overall preferences and to detect potential biases, such as an aversion to AI-generated content (ChatGPT). Each participant made two recommendations, one for each pair of answers. An overall recommendation score was determined for each participant (see Table 2).

Table 2. *Criteria for Determining Each Participant's Overall Recommendation for Answers.*

Overall recommendation	Criteria for determining overall recommendation
Recommend health professional's answers	Participants exclusively recommending one or both answers from health professionals without exclusively recommending an answer from ChatGPT.
Recommend both answers	Participants (a) recommending both answers for both answer pairs or (b) recommending an answer by ChatGPT for one answer pair and an answer by health professionals for the other.
Recommend ChatGPT answers	Participants exclusively recommending one or both answers by ChatGPT without exclusively recommending an answer by health professionals.

Qualitative Data Collection: Motivation of Answer Recommendations

To capture the reasoning behind participants' evaluations and to provide qualitative data for the study, participants responded to the following open-ended question after each answer recommendation: *Write 3–4 sentences on why you would recommend one answer over the other, or why they were equally good.* Each participant provided two such open-ended explanations.

Ethical Considerations

Participation in the study was entirely voluntary. No IP addresses or personally identifiable information were collected, ensuring participant anonymity. To minimize potential distress, we removed any questions related to medical diagnoses. The study was approved by the Norwegian Agency for Shared Services in Education and Research (Sikt).

Quantitative Data Analysis

Quantitative analyses examined how participants' evaluations and answer recommendations were influenced by knowledge of the author's identity (source knowledge) and the author type (health professional or ChatGPT). All analyses were performed using SPSS v29.

Two main types of comparisons were conducted: within-group comparisons, assessing differences between answers authored by ChatGPT and those authored by health professionals separately within each condition; and between-group comparisons, evaluating differences in participants' assessments across the different conditions.

Answer Evaluations

Differences in participants' evaluations by source knowledge and author type were analyzed using four dependent variables: Validation, Relevance, Clarity, and Utility. Within-group comparisons were conducted using paired-samples *t*-tests to compare participants' evaluations of ChatGPT with those of health professionals within each source condition (known and unknown). Between-group comparisons were conducted using independent-samples *t*-tests to compare evaluations between participants in the source-known and source-unknown conditions. The tests were conducted without corrections.

All tests were two-tailed, with no directional hypotheses proposed.

Answer Recommendations

Participants' answer recommendations were also analyzed to explore preference patterns and potential biases (e.g., AI aversion). Within-group comparisons were conducted using Chi-Square tests for Goodness of Fit to examine the proportion of participants recommending ChatGPT answers, health professional answers, or both within each study condition. Between-group comparisons were conducted using Chi-Square tests to assess whether the likelihood of recommending ChatGPT answers differed between participants with and without source knowledge.

Qualitative Data Analysis

We employed thematic analysis to systematically identify and interpret patterns of meaning within participants' qualitative responses. This approach was particularly suited to the study's aim of understanding participants' subjective reasoning for their answer preferences. More specifically, we followed the six-step procedure for thematic analysis developed by Braun and Clarke (2006). That procedure dictates the following five analysis steps before the sixth step of writing the report: read through the answers to gain an overview of the data; code all the answers (one author); merge similar codes into overarching themes; return to the raw data to ensure the themes represent the full dataset; name the themes.

To ensure quality in the analysis, a second author reviewed all codes and provided feedback. Disagreements occurred for 7% of the codes and were resolved through discussion, ensuring a consensus-driven interpretation of the data. We used a Chi-Square test of independence to assess whether there were significant differences in the distribution of themes between the two groups.

Results

In line with our mixed-methods approach, we present the quantitative results first, followed by the qualitative findings. This ordering allows us to establish overall patterns in participants' evaluations before turning to the open-ended responses, which provide deeper insight into the reasoning behind those evaluations, enabling interpretations of those patterns.

Quantitative Analysis Results

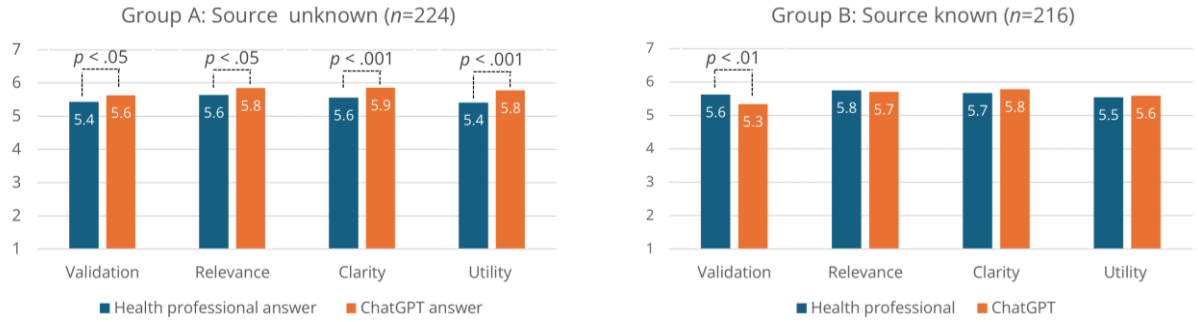
The results of the quantitative analysis are divided into two parts. First, we present the results from the within-group analysis, followed by the between-group analysis. While the between-group analysis provided insights only into the effect of Source Knowledge, the within-group analyses provided insights into the effects of Author Type and Source Knowledge.

Within-Group Analyses: The Effect of Author Type

We first investigated the within-group effects of Author Type on answer evaluation. The participants tended to rate answers from health professionals and ChatGPT above the middle scale value for Validation, Relevance, Clarity, and Utility, both when the source was unknown and when it was known. That is, the two study groups assigned average scores above 5 to both author types for each variable on the 7-point Likert scale (See Figure 2).

Paired samples *t*-tests were conducted for the four dimensions of answer evaluation for participants in the Source Unknown (n = 224) and Source Known (n = 216) conditions. The two groups differed substantially in their evaluation of answers. Participants in the Source Unknown condition tended to score answers by ChatGPT higher than answers by health professionals on all four dimensions: Validation ($t(223) = -2.08$, $p < .05$), Relevance ($t(223) = -2.56$, $p < .05$), Clarity ($t(223) = -3.73$, $p < .001$), and Utility ($t(223) = -3.97$, $p < .001$). No such effect was observed for participants in the Source Known condition, with participants instead tending to score ChatGPT answers significantly lower than those of health professionals on Validation ($t(215) = 2.77$, $p < .01$), with no significant differences observed for the other three dimensions, namely, Relevance ($t(215) = .64$, $p = .520$), Clarity ($t(215) = -1.33$, $p = .185$), and Utility ($t(215) = -.44$, $p = .659$). Figure 2 provides an overview of the findings of this analysis.

Figure 2. *Within-Group Comparison of Answer Evaluation for Health Professional Answers and ChatGPT Answers.*

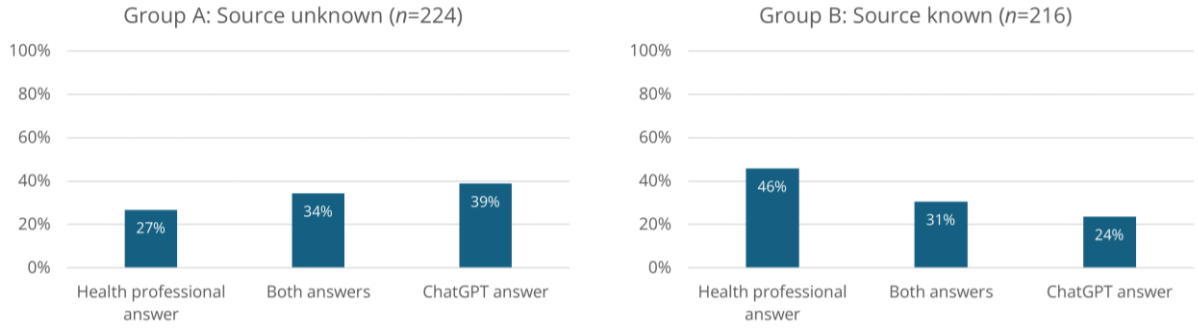


Note. Separate comparisons for the Unknown Source and Known Source conditions. Significant differences by paired samples *t*-tests.

An Author Type effect was also found for Answer Recommendation. This effect was also markedly different across conditions. For Source Unknown, more participants recommended the ChatGPT answers (39%) than those recommending both answers (34%) or the health professionals answer (27%), with a Chi-Square test for Goodness of Fit—assuming all three recommendation categories equally likely—approaching significance ($\chi^2 = 4.99$, $p = .082$).

For Source Known, significantly more participants recommended the health professionals answer (46%) than those recommending both answers (31%) or the ChatGPT answer (24%), according to a Chi-Square test for Goodness of Fit ($\chi^2 = 16.75$, $p < .001$). For a full overview of participant distribution across the recommendation categories, see Figure 3.

Figure 3. *Distribution of Answer Recommendations for Unknown Sources and Known Sources.*



Between-Groups Analyses: The Effect of Source Knowledge

To assess the implications of source knowledge for the perception of AI-generated advice, we investigate between-group effects on answer evaluation. We conducted independent samples *t*-tests comparing the Source Unknown and Source Known conditions across the four evaluation dimensions. The findings suggest that disclosing AI authorship may bias participants' evaluations of the answers to some extent. On average, answers from health personnel were rated nominally higher for all four dimensions in the Source Known condition, although no significant differences were found, with the largest nominal differences were found for Validation ($t(438) = -1.73$, $p = .085$). On average, ChatGPT answers were rated nominally higher for all four dimensions in the Source Unknown condition, although a significant difference was only observed for Validation ($t(438) = 2.44$, $p < .05$).

A significant effect of source knowledge was also observed for answer recommendation: participants in the Source Unknown condition more often recommended ChatGPT answers (39%) than those in the Source Known condition (24%), as indicated by a Chi-Square Test of Independence, $\chi^2 = 11.85$, $p < .001$.

Qualitative Analysis Results

Positive Attributes of Answers

The thematic analysis recognized that (1) cognitive, (2) relational, (3) relevance, and (4) professionalism attributes shaped positive evaluations of the answers. Figure 4 shows the distribution of these attributes, and Table 3 summarizes them by author identity.

Figure 4. *Overview of the Positive Attributes of Answers from ChatGPT and Health Professionals.*

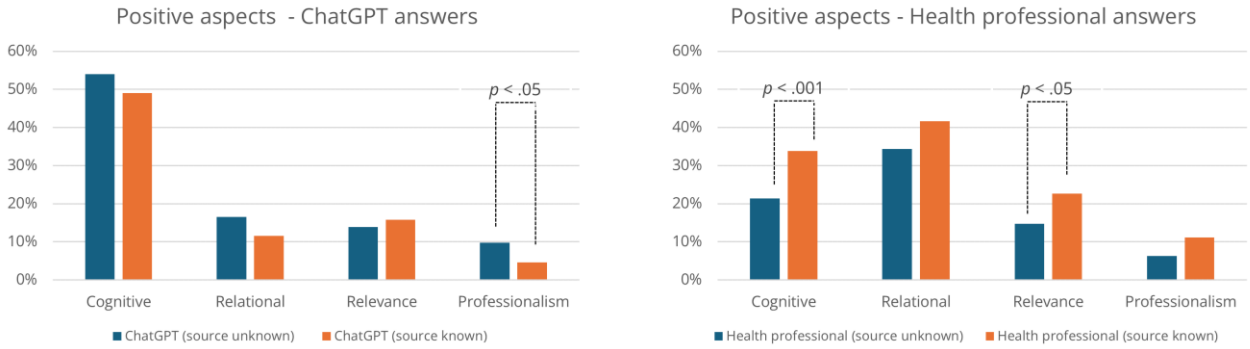


Table 3. *Overview of Themes Contributing to a Positive Evaluation of the Answer.*

Attributes	Themes	Frequency					
		ChatGPT		Health professional		Both	
		Unknown	*Known*	*Unknown*	*Known*	*Unknown*	*Known*
Cognitive	*Concrete and actionable advice*	34% (76)	22% (48)	5% (11)	16% (34)	12% (27)	11% (23)
	Understandable and easy to process	25% (55)	27% (58)	9% (20)	12% (26)	8% (19)	3% (7)
	Comprehensive and explanatory	21% (47)	15% (33)	13% (29)	14% (30)	5% (11)	4% (9)
Relational	*Compassionate and empathic tone*	12% (27)	8% (18)	30% (67)	34% (74)	5% (11)	6% (14)
	Validates or normalizes	6% (14)	4% (9)	7% (15)	6% (12)	2% (5)	3% (6)
	Humanlike	1% (2)	0% (1)	6% (13)	7% (15)	—	0% (1)
Relevance	*Answer adapted to the person asking*	14% (31)	16% (34)	15% (33)	23% (49)	3% (7)	2% (5)
Professionalism	*Professional and credible*	9% (20)	3% (6)	6% (14)	10% (22)	1% (2)	0% (1)
	Avoids medical diagnoses	1% (3)	2% (4)	0% (1)	1% (2)	—	0% (1)
Other	*Suggests help services or resources*	7% (15)	6% (14)	4% (9)	6% (12)	3% (7)	5% (10)
	Answer written by a health professional	—	—	—	24% (52)	—	—

Note. The themes of this analysis largely overlap with those of a similar study (In review). The analysis here, however, is on data from a different sample and for a different study objective (submitted).

Positive Cognitive Attributes. Participants offering positive reports reflecting cognitive qualities valued answers that they perceived as concrete and useful, often comprising tips, actionable advice, and potential solutions to the problems raised by the young people. For example, Participant ID51 (Source Unknown) wrote the following regarding a ChatGPT answer: "It provides several tips that can help if you're afraid to tell an adult about what you're experiencing. So you can try some techniques yourself before possibly reaching out."

Other participants considered it positive if an answer was detailed or provided extensive explanations of the problem, as in the following comment from Participant ID219 (Source Unknown) regarding a ChatGPT answer: "There's more information. It explains techniques. It feels like it has searched more on the internet."

Others emphasized clear text structure and ease of understanding, achieved through bullet points, simple language, and the avoidance of redundancy.

Reports reflecting positive cognitive attributes were more often found in relation to ChatGPT-generated answers (52%) than those produced by health professionals (28%). Furthermore, a Chi-Square Test of Independence showed that answers authored by health professionals were significantly more often associated with positive cognitive attributes in the Source Known condition (Figure 4).

Positive Relational Attributes. Participants whose reports positively related to relational qualities often appreciated answers to concerned responses that were perceived as genuine, kind, and personal, helping recipients feel seen and heard: "It feels friendlier. It shows more understanding of the sender's situation and feelings and can provide comfort even if there might not be a solution to the problem" (ID302; Health professional answer; Source Unknown).

Typically, positive reflections on relational attributes included answers that validated or normalized the young people's feelings or situation, reduced shame or worry, and made the recipient feel less alone. Interestingly, some participants in the Source Known condition reported enjoying the answers because they were humanlike, creating a sense that they were talking to a real person:

> "The first response seems more like a person responding to you as opposed to ChatGPT, which just lists information on the topic. This makes the first response seem more genuine, and the chance that you will take it to heart is greater." (ID337; Health professional answer; Source Known)

As Figure 4 shows, positive relational attributes were more often reflected in reports on health professionals' answers (38%) than those concerning ChatGPT's answers (14%). However, there were no significant differences between the experimental conditions in this regard.

Positive Relevance Attributes. Participants whose reports positively captured relevance qualities appreciated answers perceived as highly pertinent to the question or tailored to the specific needs and circumstances of young people: "It's more directed toward the specific person writing, seems more understandable, and the text conveys more pathos in what's written" (ID195; Health professional; Source Unknown).

This attribute was mentioned about equally in reports concerning answers from ChatGPT (15%) and health professionals (19%). A Chi-Square Test of Independence indicated significantly more mentions of these attributes for reports on the answers of health professionals from participants in the Source Known condition compared to those in the Source Unknown condition (see Figure 4).

Positive Professionalism Attributes. Positive reports related to professionalism qualities indicated that participants valued answers they perceived as professional. Some reported that the answers appeared to be written by someone knowledgeable, while others noted that the answers felt "safe":

> "Even though ChatGPT may have some errors, it has at least written the text in a more credible and structured way. ChatGPT also includes more examples and sounds like it has used many sources, while the expert says, "based on the studies I've read." (ID505; ChatGPT answer; Source Known)

Participants also reported the attribute as positive if the answer avoided focusing on or insinuating a possible diagnosis. Positive professionalism attributes were reflected about equally often in reports on answers from ChatGPT (7%) and answers from health professionals (9%). However, for ChatGPT answers, a Chi-Square Test of Independence showed that reports were more likely to capture perceived positive professionalism when the source was unknown (Figure 4).

Other Attributes. Other answers incorporated various topics, including references to external resources, relevant books, articles, and professional contacts. Some participants also reported it as positive if an answer was written by a health professional and not ChatGPT, without elaborating on why.

Negative Attributes

Cognitive, relational, relevance, and professionalism attributes also contributed to negative evaluations, mirroring the patterns observed for positive evaluations. Figure 5 illustrates the relative distribution of these attributes, and Table 4 summarizes them and their association with author identity (ChatGPT vs. human; identity known vs. unknown).

Figure 5. *Overview of Perceived Negative Attributes of Answers from ChatGPT and Health Professionals.*

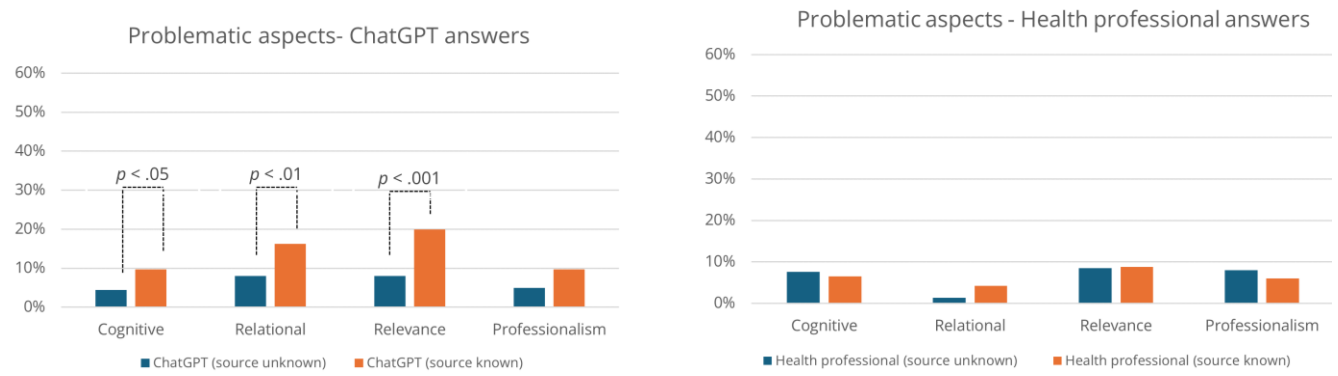


Table 4. Overview of Themes Concerning Negative Evaluations of the Answer.

Attributes	Themes	Frequency					
		ChatGPT		Health professional		Both	
		Unknown	Known	Unknown	Known	Unknown	Known
Cognitive	*Too detailed or difficult to process*	4% (9)	9% (20)	6% (13)	5% (10)	1% (2)	0% (1)
	Poorly structured text and grammatical errors	1% (2)	0% (1)	2% (4)	2% (5)	—	0% (1)
Relational	*Too factual and cold*	8% (18)	16% (35)	1% (3)	4% (9)	—	0% (1)
Relevance	*Too general*	7% (16)	19% (42)	3% (7)	3% (7)	—	—
	Lack of concrete advice and explanations	2% (4)	0% (1)	5% (12)	6% (14)	—	1% (3)
Professionalism	*Unprofessional and uncertain*	1% (3)	3% (6)	4% (9)	1% (3)	—	—
	Insisting and assertive	3% (6)	7% (15)	2% (5)	1% (2)	0% (1)	—
	Focus on medical diagnosis	1% (2)	0% (1)	3% (6)	4% (8)	—	—
Other	*Answer is or appear AI generated*	4% (8)	16% (36)	—	—	—	—

Negative Cognitive Attributes. Participants evaluated answers negatively in terms of cognitive attributes when they were difficult to understand, often due to excessive detail or overly complex information: "ChatGPT included more details but wasn't always as objective and was, in some places, harder to understand" (ID646; ChatGPT answer; Source Known).

Participants also reported poorly structured answers that appeared messy, with complex wording and grammatical errors. As shown in a Chi-Square Test of Independence, negative cognitive attributes were mentioned significantly more often for ChatGPT answers when authorship was known (Figure 5).

Negative Relational Attributes. Relational qualities in answers were experienced as problematic when they came across as too cold, factual, or clinical, lacking the expected warmth and empathy: "The ChatGPT response is far too concrete, and you don't get as close—it doesn't become a real dialogue and doesn't show any genuine empathy or understanding for the person's feelings, other than a superficial understanding" (ID85; ChatGPT answer; Source Known).

Negative relational attributes were more frequently associated with ChatGPT answers, particularly among participants in the Source Known condition. The difference between the Source Unknown and Source Known conditions was significant according to a Chi-Square Test of Independence (See Figure 5).

Negative Relevance Attributes. Answers that negatively captured relevance qualities were typically reported as lacking relevance or being too general or superficial. "The answer written by ChatGPT didn't seem like the question was properly read. It didn't really understand what the challenges were, and an AI can't relate to or empathize with those kinds of feelings" (ID85; ChatGPT answer; Source Known).

When ChatGPT was identified as the source, participants more often reported a lack of concrete advice, an observation confirmed by the Chi-Square Test of Independence (Figure 5).

Negative Professionalism Attributes. Negative professionalism qualities were documented when answers were perceived as lacking professionalism, containing excessive caveats that introduced uncertainty, or being condescending toward the young person posing the question. "The second one seemed more professional. The first one felt like there was a lot of uncertainty and guessing, while the second one came across as a professional. The answers were clear" (ID549; Health professional; Source Unknown).

Some participants evaluated answers particularly negatively when the author appeared overly insistent or assertive, or when the response introduced discussions of diagnosis: "The first thing ChatGPT does is say that what the person is experiencing could be symptoms of depression, anxiety, and other mental health issues, which doesn't really help and feels a bit strange" (ID60; ChatGPT answer; Source Known).

Negative professionalism attributes were nominally more often mentioned in relation to ChatGPT in the Source Known condition. However, the differences between the conditions were not statistically significant.

Other Attributes. Some participants evaluated ChatGPT-generated answers negatively on other grounds, indicating that an AI can never truly understand what the help-seeking person is going through, or that it can make mistakes. Thus, some recommended not using AI in this context, as in the following example: "Again, do not use ChatGPT for matters involving psychiatric help. It seems like people's health is not being valued if ChatGPT is used" (ID875; ChatGPT answer; Source Known).

Discussion

Our results provide insights into young people's perceptions of mental health advice from health professionals and ChatGPT, accounting for source awareness. As such, these findings enable a thorough discussion of our three research questions. First, we discuss young people's overall perceptions of mental health advice from the two sources (RQ1), then explore the key attributes of the advice that shape these perceptions (RQ2), and finally detail how source awareness impacted their perceptions (RQ3).

Young People's Perceptions of Mental Health Advice from LLMs vs. Health Professionals (RQ1)

Although we know from existing research and practice the benefit of mental health advice from health professionals through online support websites or helplines, and we have also seen emerging evidence on the uptake and perceived benefit of LLMs for mental health support, it is interesting to observe our participants' evaluations of the presented answers to help-seeking questions included in our study. All answers provided by ChatGPT and health professionals were assessed on four dimensions drawn from the Session Rating Scale, showing generally high agreement with the perceived Validation, Relevance, Clarity, and Utility of the provided advice. On average, answers from both ChatGPT and health professionals scored above 5 on a 7-point scale across all dimensions, regardless of source awareness. Furthermore, the qualitative reports showed that recommended answers, whether from ChatGPT or health participants, were valued for positive cognitive and relational attributes, as well as relevance and professionalism. Furthermore, references to the positive attributes of both human-provided and AI-generated answers far outnumber references to the negative attributes. Hence, participants seemed to find value and support in the answers provided, regardless of the source. This may suggest that LLMs can work as a source of mental health advice complementary to that of health professionals.

This implication is not unexpected. Instead, existing research has indeed suggested the potential benefits of LLMs as a source of mental health support (Khurana et al., 2024). In particular, LLMs have shown promise in alleviating mental health concerns (Hua et al., 2024; Jiang et al., 2024; T. Kim et al., 2024; Na, 2024). Research has also suggested a positive inclination among people, in general (Mendel et al., 2025), and among young people, in particular (Brandtzæg et al., 2021; Brandtzaeg et al., 2025), to take up AI-powered support for mental health. Furthermore, previous work has shown that AI-generated advice may have preferable qualities in health contexts (e.g., display of empathy and detailed responses; Ayers et al., 2023; J. Kim et al., 2024).

As such, our findings in response to our first research question serve, in part, to confirm indications from previous work and, in part, to indicate the potential benefits offered by both sources of health advice. As such, the question that emerges concerns not so much whether LLM-generated advice can be perceived as beneficial by help-seeking users, but rather how and for which contexts. It will be important to understand the perceived strengths and weaknesses of such advice, as well as how source disclosure impacts these perceptions.

Key Attributes Shaping Young People's Evaluations of Mental Health Advice From ChatGPT and Health Professionals (RQ2)

Our qualitative findings revealed that evaluations of mental health advice were shaped by four key attributes: cognitive quality, relational warmth, relevance of the answer, and professionalism. Thus, answers perceived as clear, empathetic, situation-relevant, and professionally competent were evaluated most positively. ChatGPT's answers were often described as action-oriented, easy to understand, practical, and empathetic. However, participants also noted instances where ChatGPT's answers lacked professionalism or relevance or exhibited an overly factual tone.

These results align with the existing literature on human–AI interactions, which consistently indicates a preference for advice authored by LLMs, such as ChatGPT, when authorship is concealed. This preference seems to be related to LLMs' user-friendly writing and formatting style (Herbold et al., 2023; Singhal et al., 2025), their ability to provide detailed answers (Young et al., 2024), and their empathic tone (Ayers et al., 2023; Vowels, 2024). In line with Young et al. (2024), our study demonstrates that in a youth-focused mental health help-seeking context, these qualities are important, and further suggests that LLMs may serve a purpose in this context.

While we will go into greater depth on the implications of source awareness, it can already be useful at this point in the discussion to detail how this impacted qualitative assessments. Negative attributes of ChatGPT-generated answers were most often reported when participants knew the author's identity, suggesting that source awareness influenced the perception of the answers. Consistent with these patterns, answers written by health professionals were reported as more caring and better tailored when their human identity was known.

Some participants also explicitly expressed discomfort with ChatGPT's answers, indicating that AI aversion persists in specific contexts, particularly those involving emotional sensitivity, trust, or care. Interestingly, when the source was not disclosed, participants often preferred the content generated by ChatGPT, suggesting that, in the absence of identity cues, generative AI can outperform human experts in perceived quality.

How Source Awareness Influences Young People's Mental Health Advice Preferences (RQ3)

Our results reveal the critical role of source awareness in shaping young people's perceptions of AI-generated content in the context of mental health and help-seeking. Significantly more participants in the Source Unknown condition preferred ChatGPT's answers compared to participants in the Source Known condition. Furthermore, source knowledge impacted participant assessments of the provided mental health advice. These quantitative results show that perceptions of AI-generated advice are shaped not only by the content of the answers but also by users' perceptions of the source. Furthermore, as noted above, source awareness was also found to impact qualitative assessments.

The differences observed between participants in the two conditions may be attributed to AI aversion, particularly regarding AI's perceived inability to exhibit what many see as typical "human" qualities, such as validation, empathy, and professionalism. Our findings indicate that disclosing that an answer is AI-generated may activate pre-existing beliefs, biases, and expectations about what AI can or cannot do, especially in emotionally sensitive contexts such as mental health. As Vowels (2024) has documented, people often associate AI, such as LLMs, with being impersonal, mechanical, or lacking empathy, which can lead them to re-evaluate the same content more negatively once they know it is authored by an AI. Hence, capturing the notion of AI aversion, mental health advice may be assessed not only in terms of its content but also in terms of its source.

The sensitivity associated with mental health help-seeking may amplify AI aversion due to the emotionally complex nature of such interactions. The extant literature has similarly indicated greater aversion towards AI systems performing subjective tasks, such as providing dating advice (Castelo et al., 2019). Interestingly, it has been proposed that enhancing human-likeness can mitigate such aversions (Castelo et al., 2019). Although our study did not explicitly measure human-likeness, LLMs such as ChatGPT are generally perceived as highly humanlike due to their advanced conversational abilities (Sorin et al., 2024).

Moreover, AI aversion may be stronger if users perceive algorithmic mistakes (Jussupow et al., 2020). Although none of our participants reported errors in ChatGPT's answers, it is likely that the public discourse surrounding LLMs and the potential for hallucinations may impact this. Such potential aversive attitudes toward AI should also

be contrasted with the previously documented benefits of AI-based systems for mental health support, including perceived anonymity and the absence of social stigma and feelings of judgment (Brandtzæg et al., 2021).

Yet, it has been argued in other contexts that such reactions may not stem from AI aversion alone but also reflect a form of human favoritism (Zhang & Gosline, 2023), a cognitive bias that sees people overvalue content when they believe it was produced by a human, especially an expert, relative to when the same content is believed to be produced by AI. If AI aversion were the sole driver of our study findings, we would expect differences only when answers were labeled as AI-generated. However, our findings show that participants rated answers more positively—especially in terms of empathy and person-orientation—when they believed they were provided by human health professionals. This suggests a combined effect of AI aversion and human favoritism, pointing to a more complex bias dynamic that may be rooted less in concerns about competence and more in violations of social expectations regarding care and help-seeking.

Finally, the growing presence of AI in everyday life among young people may lead to its normalization, potentially reducing aversion over time as new social expectations about what AI can do take hold. In our study, participants reported varying levels of exposure to different generative AI tools. As people become more familiar with and comfortable using LLMs and AI systems, initial skepticism may give way to broader acceptance, also in the context of help-seeking.

Practical and Theoretical Contribution

This study contributes to the emerging field of human–AI interaction by demonstrating that source attribution significantly influences the reception of AI-generated content in sensitive domains such as mental health support and help-seeking.

Our findings reveal that young people's judgments about AI-generated mental health advice are shaped by their awareness of the author's identity and the specific qualities of the answers themselves, namely, cognitive, relational, relevance, and professionalism attributes. This insight advances theoretical understanding of how content-level attributes mediate perceptions of AI credibility and usefulness, extending human–AI interaction research into emotionally sensitive contexts, such as mental health. The observed shift in preferences in response to source disclosure raises critical questions about transparency and user engagement.

Here, we argue that although AI aversion in the short term may reduce the uptake of potentially beneficial AI applications in mental health support, it may also serve as a barrier, helping users and providers avoid overreliance. This may be particularly important in the early phase of AI-powered health support, when quality criteria and processes may not be sufficiently established. Furthermore, for service providers, it may be important to be aware of AI aversion and human favoritism when designing health support services that align with users' evolving preferences.

Based on the above, a promising avenue for current mental health support services is likely a hybrid approach in which LLM-powered support is integrated into a service system that sees health professionals and LLMs complement each other in a transparent manner to serve users effectively and efficiently. Specifically, our findings suggest that LLMs, if carefully integrated, could serve as valuable preliminary resources in youth mental health support services, offering clear, empathetic, and practical advice when human counselors are not immediately available or for questions that may not require the attention of health personnel.

Limitations

Although our study provides valuable insights into human–AI interaction in general and AI aversion in the context of mental health help-seeking in particular, several limitations warrant further investigation. Using a stratified sample would enhance generalizability within the Norwegian context, particularly given the country's cultural specificity, such as high levels of trust in public institutions, which may limit the study's wider applicability.

Although the questions used in this study were generated by real young people seeking help for genuine mental health concerns, the participants evaluating the answers were not the original questioners. This creates an artificial evaluation context that may not fully reflect real-world help-seeking interactions. However, because the questions addressed authentic issues relevant to many young people, the study remains relevant to real-world help-seeking situations.

Importantly, we examined only young people's perceptions of answers to less sensitive mental health questions. Additionally, our study focused solely on how young people perceived the answers. That is, we did not assess the quality of either the answers generated by ChatGPT or those provided by health professionals.

It is also important to note that health professionals at *ung.no* are subject to strict guidelines when responding to questions about youth posts on their platform. They are not allowed to provide healthcare recommendations, which are defined as "any action that has a preventive, diagnostic, therapeutic, health-preserving, rehabilitative, or nursing and care purpose and is carried out by health personnel."[2] These guidelines may have impacted how helpful the answers were perceived to be. Conversely, ChatGPT may not adhere to such guidelines and may have produced answers that appeared more helpful.

Moreover, we only used ChatGPT to generate answers and participants were informed that this specific LLM had produced the responses. Seen as ChatGPT was a widely used LLM at the time of the study, it may have impacted on the participants' perceptions. Future studies should explore whether AI aversion may change depending on the type of LLM involved.

Future research—which is already planned—should explore the phenomenon through a longitudinal design to investigate whether AI aversion among young people in the context of mental health and help-seeking diminishes over time, as exposure to and familiarity with AI increase.

Footnotes

[1] www.limesurvey.org

[2] https://www.npe.no/en/contribution-scheme/limitations-to-the-registration-duty/what-is-considered-healthcare-under-the-patient-injury-act/

[i] https://www.ipsos.com/nb-no/datainnsamling-leveranse

Conflict of Interest

The authors have no conflicts of interest to declare.

Use of AI Services

During the preparation of this work, the authors used ChatGPT to improve language and readability. Generative AI and AI-assisted technologies were not used to support other parts of the writing process. Following the use of this tool, the authors thoroughly reviewed and edited the content as needed and take full responsibility for the final content of the publication.

Authors' Contribution

Petter Bae Brandtzaeg: conceptualization, methodology, funding acquisition, supervision, project administration, validation, writing—original draft, writing—review & editing. **Marita Skjuve:** conceptualization, data curation, investigation, methodology, formal analysis, visualization, writing—original draft, writing—review & editing. **Asbjørn Følstad:** conceptualization, formal analysis, methodology, resources, supervision, validation, writing—review & editing.

Acknowledgement

The study was funded by the Dam Foundation. Grant number SDAM_FOR462134.

References

Arvai, N., Katonai, G., & Meskó, B. (2025). Health care professionals' concerns about medical AI and psychological barriers and strategies for successful implementation: Scoping review. *Journal of Medical Internet Research*, *27*, Article e66986. https://doi.org/10.2196/66986

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, *183*(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

Bickmore, T., Gruber, A., & Picard, R. (2005). Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient Education and Counseling*, *59*(1), 21–30. https://doi.org/10.1016/j.pec.2004.09.008

Brandtzæg, P. B. B., Skjuve, M., Dysthe, K. K. K., & Følstad, A. (2021). When the social becomes non-human: Young people's perception of social support in chatbots. In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445318

Brandtzaeg, B. P., Skjuve, M., & Følstad, A. (2025). AI individualism: Transforming social structures in the age of social artificial intelligence. In P. Hacker (Ed.), *Oxford Intersections: AI in Society.* Oxford University Press. https://doi.org/10.1093/9780198945215.003.0099

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G. M., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (NeurIPS 2020). Curran Associates.

Böhm, R., Jörling, M., Reiter, L., & Fuchs, C. (2023). People devalue generative AI's competence but not its advice in addressing societal and personal challenges. *Communications Psychology*, *1*, Article 32. https://doi.org/10.1038/s44271-023-00032-x

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Dang, J., & Liu, L. (2024). Extended artificial intelligence aversion: People deny humanness to artificial intelligence users. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspi0000480

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376638

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Duncan, B. L., Miller, S. D., Sparks, J. A., Claud, D. A., Reynolds, L. R., Brown, J., & Johnson, L. D. (2003). The session rating scale: Preliminary psychometric properties of a "working" alliance measure. *Journal of Brief Therapy*, *3*(1), 3–12. https://betteroutcomesnow.com/wp-content/uploads/session-rating-scale.pdf

Durairaj, K. K., Baker, O., Bertossi, D., Dayan, S., Karimi, K., Kim, R., Most, S., Robotti, E., & Rosengaus, F. (2024). Artificial intelligence versus expert plastic surgeon: Comparative study shows ChatGPT "wins" rhinoplasty consultations: Should we be worried? *Facial Plastic Surgery & Aesthetic Medicine*, *26*(3), 270–275. https://doi.org/10.1089/fpsam.2023.0224

Gaskin, J., E. Blondeel, T. Bullock, R. Schuetzler, R. Serre, J. Steffen., & D. A. Wood. (2024). Chatbots mitigate help-seeking avoidance. *International Conference on Information Systems*, *45*, 1–15. https://scholarsarchive.byu.edu/facpub/8252

Hatch, S. G., Goodman, Z. T., Vowels, L., Hatch, H. D., Brown, A. L., Guttman, S., Le, Y., Bailey, B., Bailey, R. J., Esplin, C. R., Harris, S. M., Holt, D. P., McLaughlin, M., O'Connell, P., Rothman, K., Ritchie, L., Top, D. N., & Braithwaite, S. R. (2025). When ELIZA meets therapists: A Turing test for the heart and mind. *PLoS Mental Health*, *2*(2), Article e0000145. https://doi.org/10.1371/journal.pmen.0000145

Henestrosa, A. L., & Kimmerle, J. (2024). The effects of assumed AI vs. human authorship on the perception of a GPT-generated text. *Journalism and Media*, *5*(3), 1085–1097. https://doi.org/10.3390/journalmedia5030069

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, *13*, Article 18617. https://doi.org/10.1038/s41598-023-45644-9

Hua, Y., Na, H., Li, Z., Liu, F., Fang, X., Clifton, D., & Torous, J. (2024). Applying and evaluating large language models in mental health care: A scoping review of human-assessed generative tasks. *arXiv*. https://doi.org/10.48550/arxiv.2408.11288

Hudson, A. L., Nyamathi, A., Greengold, B., Slagle, A., Koniak-Griffin, D., Khalilifard, F., & Getzoff, D. (2010). Health-seeking challenges among homeless youth. *Nursing Research*, *59*(3), 212–218. https://doi.org/10.1097/nnr.0b013e3181d1a8a9

Jiang, M., Zhao, Q., Li, J., Wang, F., He, T., Cheng, X., Yang, B. X., Ho, G. W. K., & Fu, G. (2024). A generic review of integrating artificial intelligence in cognitive behavioral therapy. *arXiv*. https://doi.org/10.48550/arxiv.2407.19422

Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *Npj Digital Medicine*, *6*, Article 6. https://doi.org/10.1038/s41746-023-00751-9

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In *Proceedings of the 28th European Conference on Information Systems (ECIS 2020).* ECIS. https://aisel.aisnet.org/ecis2020_rp/168

Kacperski, C., Ulloa, R., Bonnay, D., Kulshrestha, J., Selb, P., & Spitz, A. (2025). Characteristics of ChatGPT users from Germany: Implications for the digital divide from web tracking data. *PLoS ONE*, *20*(1), Article e0309047. https://doi.org/10.1371/journal.pone.0309047

Khurana, A., Subramonyam, H., & Chilana, P. K. (2024). Why and when LLM-based assistants can go wrong: Investigating the effectiveness of prompt-based interactions for software help-seeking. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (pp. 288–303). Association for Computing Machinery. https://doi.org/10.1145/3640543.3645200

Kim, J., Lee, S.-Y., Kim, J. H., Shin, D.-H., Oh, E. H., Kim, J. A., & Cho, J. W. (2024). ChatGPT vs. sleep disorder specialist responses to common sleep queries: Ratings by experts and laypeople. *Sleep Health*, *10*(6), 665–670. https://doi.org/10.1016/j.sleh.2024.08.011

Kim, T., Bae, S., Kim, H. A., Lee, S.-W., Hong, H., Yang, C., & Kim, Y.-H. (2024). MindfulDiary: Harnessing large language model to support psychiatric patients' journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–20). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642937

Mendel, T., Singh, N., Mann, D. M., Wiesenfeld, B., & Nov, O. (2025). Laypeople's use of and attitudes toward large language models and search engines for health queries: Survey study. *Journal of Medical Internet Research*, *27*, Article e64290. https://doi.org/10.2196/64290

Na, H. (2024). CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. *arXiv*. https://doi.org/10.48550/arxiv.2403.16008

Osborne, M. R., & Bailey, E. R. (2025). Me vs. the machine? Subjective evaluations of human- and AI-generated advice. *Scientific Reports*, *15*, Article 3980. https://doi.org/10.1038/s41598-025-86623-6

Pretorius, C., Chambers, D., & Coyle, D. (2019). Young people's online help-seeking and mental health difficulties: Systematic narrative review. *Journal of Medical Internet Research*, *21*(11), Article e13873. https://doi.org/10.2196/13873

Proksch, S., Schühle, J., Streeb, E., Weymann, F., Luther, T., & Kimmerle, J. (2024). The impact of text topic and assumed human vs. AI authorship on competence and quality assessment. *Frontiers in Artificial Intelligence*, *7*, Article 1412710. https://doi.org/10.3389/frai.2024.1412710

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., Neal, D., Rashid, Q. M., Schaekermann, M., Wang, A., Dash, D., Chen, J. H., Shah, N. H., Lachgar, S., Mansfield, P. A., . . . Natarajan, V. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, *31*(3), 943–950. https://doi.org/10.1038/s41591-024-03423-7

Small, W. R., Wiesenfeld, B., Brandfield-Harvey, B., Jonassen, Z., Mandal, S., Stevens, E. R., Major, V. J., Lostraglio, E., Szerencsy, A., Jones, S., Aphinyanaphongs, Y., Johnson, S. B., Nov, O., & Mann, D. (2024). Large language model–based responses to patients' in-basket messages. *JAMA Network Open*, *7*(7), Article e2422399. https://doi.org/10.1001/jamanetworkopen.2024.22399

Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., & Klang, E. (2024). Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, *26*, Article e52597. https://doi.org/10.2196/52597

Simpson, K. S. W., Gallagher, A., Ronis, S. T., Miller, D. A. A., & Tilleczek, K. C. (2022). Youths' perceived impact of invalidation and validation on their mental health treatment journeys. *Administration and Policy in Mental Health and Mental Health Services Research*, *49*(3), 476–489. https://doi.org/10.1007/s10488-021-01177-9

Vowels, L. M. (2024). Are chatbots the new relationship experts? Insights from three studies. *Computers in Human Behavior: Artificial Humans*, *2*(2), Article 100077. https://doi.org/10.1016/j.chbah.2024.100077

Young, J., Jawara, L. M., Nguyen, D. N., Daly, B., Huh-Yoo, J., & Razi, A. (2024). The role of AI in peer support for young people: A study of preferences for human-and AI-generated responses. In *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–18). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642574

Zhang, Y., & Gosline, R. (2023). Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making*, *18*, Article e41. https://doi.org/10.1017/jdm.2023.37

About Authors

Petter Bae Brandtzaeg is a Professor of Media and Communication at the University of Oslo and a Chief Scientist at SINTEF. His research focuses on human–AI interactions from a user perspective and the broader social consequences of AI. He has published in journals such as *New Media & Society*, *Human Communication Research*, and *Communications of the ACM*. Petter is the project manager of the project "Enhancing Digital Support Services for Youth through Artificial Intelligence."

https://orcid.org/0000-0002-9010-0800

Marita Skjuve is a research scientist at SINTEF and recently completed a PhD in psychology on human–AI relationships. Her work focuses on human–AI interaction in areas such as health and customer service, as well as in the private sphere. She has published in leading journals in human–computer interaction and psychology.

https://orcid.org/0000-0002-1316-9951

Asbjørn Følstad is Chief Scientist at SINTEF, a Norwegian independent research organization, with a PhD in psychology from the University of Oslo. With basis in human-computer interaction, his recent research addresses conversational user interfaces and human-centred AI. He is particularly interested in how AI is understood and taken up by individuals, organizations, and society, and how to design AI-based solutions for good user experiences.

https://orcid.org/0000-0003-2763-0996

✉ **Correspondence to**
Petter Bae Brandtzaeg, SINTEF Digital, Forskningsveien 1, 0373 Oslo, Norway and Department of Media and Communication, University of Oslo, Blindern, P.O. Box 1093, 0317, Oslo, Norway, p.b.brandtzag@media.uio.no