PeerCoPilot: A Language Model-Powered Assistant for Behavioral Health Organizations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Behavioral health conditions, which include mental health and substance use disorders, are the leading disease burden in the United States. Peer-run behavioral health organizations (PROs) critically assist individuals facing these conditions by combining mental health services with assistance for needs such as income, employment, and housing. However, limited funds and staffing make it difficult for PROs to address all service user needs. To assist peer providers at PROs, we introduce **PEERCOPILOT**, a large language model (LLM)-powered assistant that helps peer providers create wellness plans, construct step-by-step goals, and locate organizational resources. PEERCOPILOT ensures information reliability through a retrieval-augmented generation pipeline backed by a large database of over 1,300 vetted resources. We conducted human evaluations with 15 peer providers and 6 service users and found that over 90% of users supported using PEERCOPILOT. Moreover, we demonstrate that PEERCOPILOT provides more reliable and specific information than a baseline LLM. PEERCOPILOT is now used by a group of peer providers at a large behavioral health organization serving between 15 and 100 individuals daily through its network of 15 community wellness.

1 Introduction

Behavioral health conditions, including mental health and substance use disorders, are the leading disease burden in the United States, costing over $80 billion annually [Kamal et al., 2017]. Peer-run behavioral health organizations, referred to as PROs, address this critical issue in difficult-to-engage communities facing disproportionately high rates of poverty, unemployment, and housing instability [Kadakia et al., 2022, Correll et al., 2022]. PROs tackle these issues through peer providers, who leverage their personal behavioral health experiences to provide service users with wellness support and resources for housing, financial, and employment resources [Ostrow and Hayes, 2015]. PROs are transformative for individuals with behavioral conditions.

While service user demands grow year-to-year, PRO capacity has not kept up, leading to overburdened peer providers [Wall et al., 2022]. Increases in the prevalence of substance use and mental health disorders have led to growing service user demands [Counts and Nuzum, 2022]. At the same time, many PROs are underfunded, limiting their ability to train and hire new peer providers, which becomes especially pressing due to high burnout rates for peer providers [Ostrow and Leaf, 2014]. Adding to this burden is an underdeveloped technological infrastructure, with many PROs lacking systems even to manage tracking the type and range of wellness supports they provide to people they serve.

To support PROs, we propose using large language models (LLMs) as an assistant to peer providers. LLMs have had success in other domains as a tool for information retrieval [Wang et al., 2024a, Agarwal et al., 2024, Liang et al., 2024]. As a result, LLMs present a promising opportunity for PROs by potentially helping peer providers craft tailored wellness plans and synthesize location-specific

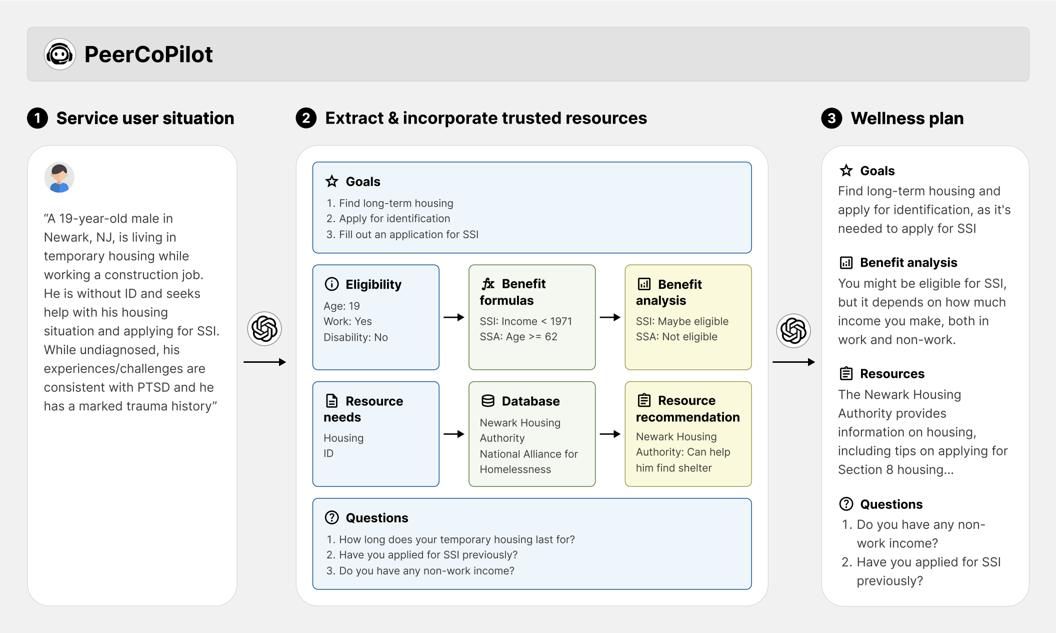

PeerCoPilot
1 Service user situation
"A 19-year-old male in Newark, NJ, is living in temporary housing while working a construction job. He is without ID and seeks help with his housing situation and applying for SSI. While undiagnosed, his experiences/challenges are consistent with PTSD and he has a marked trauma history"
2 Extract & incorporate trusted resources
Goals
1. Find long-term housing
2. Apply for identification
3. Fill out an application for SSI
Eligibility
Age: 19
Work: Yes
Disability: No
Benefit formulas
SSI: Income < 1971
SSA: Age >= 62
Benefit analysis
SSI: Maybe eligible
SSA: Not eligible
Resource needs
Housing
ID
Database
Newark Housing Authority
National Alliance for Homelessness
Resource recommendation
Newark Housing Authority: Can help him find shelter
Questions
1. How long does your temporary housing last for?
2. Have you applied for SSI previously?
3. Do you have any non-work income?
3 Wellness plan
Goals
Find long-term housing and apply for identification, as it's needed to apply for SSI
Benefit analysis
You might be eligible for SSI, but it depends on how much income you make, both in work and non-work.
Resources
The Newark Housing Authority provides information on housing, including tips on applying for Section 8 housing...
Questions
1. Do you have any non-work income?
2. Have you applied for SSI previously?

Figure 1: PeerCoPilot takes in a peer provider's input and passes this into four modules: resource recommendation, benefit eligibility, goal construction, and question generation. Some modules combine the peer provider's query with externally verified information to ensure accuracy. Modules are combined to form a response that tackles goals, resources, and follow-up questions.

resources. By assisting peer providers, LLMs can increase PRO capacity and service user capacity. At the same time, while LLM-based assistants are prevalent [Liang et al., 2024], LLMs are rarely, if ever, used in PROs because Many peer providers and service users have little familiarity with LLMs. These challenges necessitate a human-centered development process when introducing LLMs.

In this paper, we introduce **PeerCoPilot**, an LLM-based tool that assists peer providers with crafting wellness plans and retrieving resources (see Figure 1). We developed PeerCoPilot in partnership with a leading PRO. After conversations with the PRO, we designed PeerCoPilot to assist with tasks such as wellness plan creation, goal construction, resource recommendation, and benefit navigation. Peer providers stressed the need for reliable information when working with LLMs, so PeerCoPilot ensures information reliability by combining an LLM-based backend with trusted resources through techniques such as retrieval augmented generation (RAG). We evaluated PeerCoPilot through 3 onsite demos, 2 annotation sessions, and 1 semi-structured interview, totaling 15 peer providers and 6 service users. Through our onsite demos, we show that peer providers and service users are willing to use PeerCoPilot, and through our annotation sessions, we find that PeerCoPilot provides reliable and specific information. Our work is deployed to a group of peer providers at our partner PRO who use PeerCoPilot in their daily operations. [1]

2 Related Works

LLMs to Support Behavioral Health Professionals While there has been little work on LLMs in a behavioral health context, LLMs have seen great success in a variety of related fields, including education [Liu et al., 2024a, Rouzegar and Makrehchi, 2024, Rodriguez et al., 2019], social science [Mou et al., 2024, Ye et al., 2024, Ziems et al., 2024], and mental health [Lai et al., 2023, Liu et al., 2023, Beredo and Ong, 2022, Crasto et al., 2021]. Most related is work in mental health that develops tools to assist professional psychologists. For example, LLMs can simulate patients via cognitive models [Wang et al., 2024b] and roleplaying [Louie et al., 2024]. Unlike traditional mental health applications, which are often clinically focused, PROs emphasize holistic wellness through housing, employment, and financial stability alongside behavioral health.

[1] We will release code and datasets after camera ready; we keep it private for now to maintain anonymity.

Copilot Tools LLM-based copilot tools are used in domains such as software [Pudari and Ernst, 2023, Jaworski and Piotrkowski, 2023], retail [Furmakiewicz et al., 2024], and health [Ren et al., 2024]. Copilot tools improve productivity by providing templates that scaffold development [Ziegler et al., 2024]. However, copilot tools could reduce critical thinking skills and induce dependence [Lee et al., 2025]. In light of this, we develop PEERCOPILOT as a way to provide peer providers with extra resources, thereby augmenting rather than replacing them.

3 Background

To assist peer providers at PROs, we develop PEERCOPILOT, an LLM-based tool for assisting peer providers during sessions with service users. We developed PEERCOPILOT based on conversations with peer providers and service users. Both groups value factually reliable and specific information. Reliability is key because incorrect information can negatively impact service users. For example, LLM-based tools could provide inaccurate information on benefit eligibility for government programs such as Medicare, which could mislead service users. Meanwhile, specific information allows peer providers to provide tailored information and generate a detailed step-by-step plan.

4 System Design

PEERCOPILOT combines an LLM backend with modules that rely on verified information sources to ensure reliability. We describe the overall frontend and backend then individual modules.

4.1 Frontend

PEERCOPILOT consists of a chat-based frontend with reset and save buttons along with a tutorial. The reset session button allows peer providers to clear the current session between peer sessions, while the save session history button allows for a written record if peer providers want to reference a session later. The tutorial button plays a three-minute video on how to use PEERCOPILOT.

4.2 Backend Structure

After receiving a peer provider's input, PEERCOPILOT crafts a response by aggregating information from modules. PEERCOPILOT relies on four modules: resource recommendation, benefit eligibility, goal construction, and question generation. After receiving outputs from all modules, PEERCOPILOT queries GPT-4 to craft a response using the information from each module. We instruct PEERCOPILOT to construct a response that holistically addresses the service user's situation following the eight dimensions of wellness framework used at our partner PRO to guide peer providers [Swarbrick, 2006].

4.3 Backend Modules

4.3.1 Resource Recommendation

The resource recommendation module combines a resource database with RAG to ensure information reliability. Our database has over 1300 resources vetted by peer providers. Given a service user's background and goals, we use GPT-4 to extract resource needs, then match these with resource descriptions in the database via RAG [Lewis et al., 2020]. RAG matches embeddings for resource needs with database entries. We construct embeddings using a SentenceTransformer with the MPNet v2 model [Song et al., 2020], and retrieve according to the L2 metric.

4.3.2 Benefit Navigation

Government benefits, such as Supplementary Social Income (SSI) and Medicaid, involve complex inclusion criteria, making it difficult to determine eligibility. Relying on GPT-4 for benefit eligibility can result in outdated or incorrect information. Instead, we first use GPT-4 to extract demographic information such as age, monthly income, and total savings. We then pass this to formulas that assess eligibility given demographic information. These formulas are manually translated from eligibility information on government websites (e.g., Administration et al. [2024]). This results in an assessment of whether a service user is likely to be eligible for each benefit.

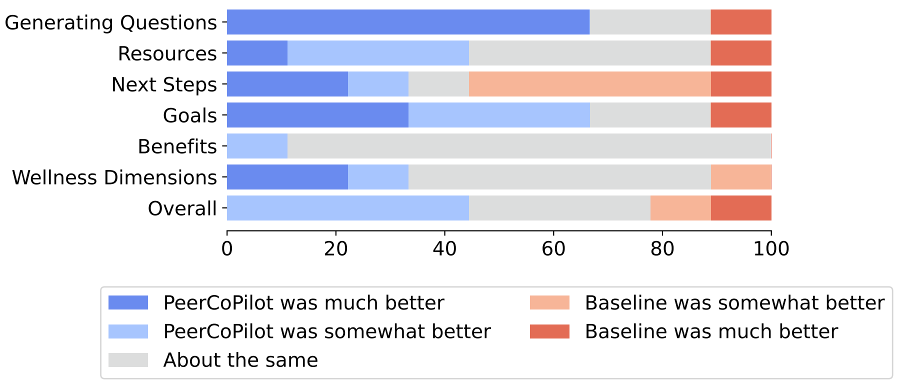


Figure 2: Peer providers find that PEERCOPILOT generates better questions, recommends better resources, and crafts better goals compared to a baseline. This is because the modules ensure information reliability and specificity.

4.3.3 Goal Construction & Question Generation

Peer providers need to offer support and construct plans tailored to service user situations. To assist with this, the goal construction module presents immediate goals for the service user, broken down into actionable steps. We construct goals by prompting GPT-4 with the SMART (Specific, Measurable, Achievable, Realistic, and Timely) goals framework [Doran, 1981], as recommended by our partner PRO. The question generation module suggests follow-up questions for peer providers to ask service users. It does so by prompting GPT-4 to craft follow-up questions prompted with the dimensions of wellness [Swarbrick, 2006] to ensure follow-up questions are holistic. Examples include “Do you have a stable place to stay?” and “Do you have transportation?”

5 PEERCOPILOT Evaluation

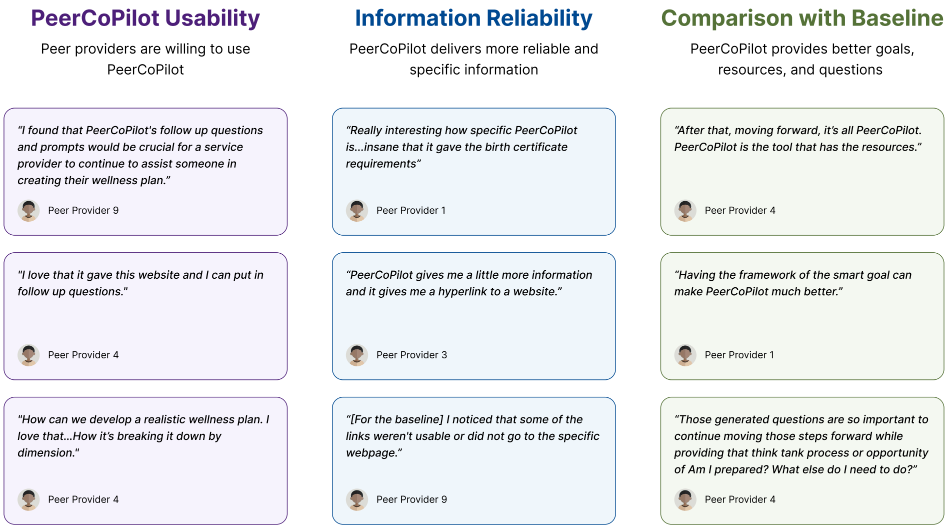
We evaluated PEERCOPILOT through human studies with peer providers and service users and found that both would use PEERCOPILOT in peer support sessions. We also find that PEERCOPILOT delivers more reliable and specific information than a GPT-4 baseline.

5.1 On-site Human Evaluations

To assess PEERCOPILOT, we conducted an on-site study with nine peer providers and six service users. Participants interact with PEERCOPILOT and baseline GPT-4 in random order to explore scenarios. We constructed nine scenarios capturing different types of situations faced by service users; we further detail these scenarios in Appendix A Participants then completed two surveys: one for system usability and another to compare PEERCOPILOT and the baseline (details in Appendix A). We include service user results in Appendix B. While we focus on results comparing against baseline GPT-4o mini, in Appendix E, we detail compare against LLMs via the LLM-as-judge framework.

Peer providers and serivce users are willing to use PEERCOPILOT We find that peer providers are willing to use PEERCOPILOT in practice. Peer providers find PEERCOPILOT simple to use, and 8 out of 9 peer providers believe that PEERCOPILOT delivers useful information for their queries. One peer provider remarks “*how we can develop a realistic plan. I love that...how it’s breaking it down by dimension.*” We further detail these results, along with results for service users, in Appendix B

PEERCOPILOT delivers more reliable and specific resources In Figure 2, we show that 4 out of 9 peer providers believe PEERCOPILOT delivers better resources, while only 1 out of 9 believe the baseline does. PEERCOPILOT delivers more reliable and specific information because it builds on top of a trusted database. Peer providers notice this difference, as one remarks “*PeerCoPilot gives me a little more information and gives me a hyperlink to a website.*” Peer providers also found PEERCOPILOT specific, with one noting that it was “*really interesting how specific PeerCoPilot*



PeerCoPilot Usability

Peer providers are willing to use PeerCoPilot

"I found that PeerCoPilot's follow up questions and prompts would be crucial for a service provider to continue to assist someone in creating their wellness plan."

Peer Provider 9

"I love that it gave this website and I can put in follow up questions."

Peer Provider 4

"How can we develop a realistic wellness plan. I love that...How it's breaking it down by dimension."

Peer Provider 4

Information Reliability

PeerCoPilot delivers more reliable and specific information

"Really interesting how specific PeerCoPilot is...insane that it gave the birth certificate requirements"

Peer Provider 1

"PeerCoPilot gives me a little more information and it gives me a hyperlink to a website."

Peer Provider 3

"[For the baseline] I noticed that some of the links weren't usable or did not go to the specific webpage."

Peer Provider 9

Comparison with Baseline

PeerCoPilot provides better goals, resources, and questions

"After that, moving forward, it's all PeerCoPilot. PeerCoPilot is the tool that has the resources."

Peer Provider 4

"Having the framework of the smart goal can make PeerCoPilot much better."

Peer Provider 1

"Those generated questions are so important to continue moving those steps forward while providing that think tank process or opportunity of Am I prepared? What else do I need to do?"

Peer Provider 4

Figure 3: We outline three themes from our sessions with peer providers: 1) PeerCoPilot provides useful information and peer providers are willing to use it, 2) PeerCoPilot provides reliable and specific information, and 3) PeerCoPilot provides better goals and questions than the baseline.

is...insane that it gave the birth certificate requirements." Conversely, for the baseline, one peer provider "*noticed that some of the links weren't usable or did not go to the specific webpage.*"

6 out of 9 peer providers prefer PeerCoPilot for goal construction and question generation (Figure 2) Peer providers found PeerCoPilot's SMART goals framework useful; one peer provider remarks "*Having the framework of the SMART goal can make PeerCoPilot much better*" because it "*spells [the goal] out.*" Peer providers also liked PeerCoPilot's follow-up questions: "*Those generated questions are so important to continue moving those steps forward while providing that think tank process or opportunity of Am I prepared? What else do I need to do?*"

Further Studies In addition to these experiments, we run two more user studies in Appendix C and Appendix D. In Appendix C, we detail a study, where we evaluated the reliability and specificity of resources suggested by PeerCoPilot. We found that PeerCoPilot delivers more reliable and specific resources than the baseline. In Appendix D, we piloted a human-AI teaming study to assess PeerCoPilot's ability in assisting peer providers with wellness plans. We show PeerCoPilot reduced completion time by 10% while leading to wellness plans better tailored to service users.

6 Discussion and Conclusion

PROs experience staffing shortages and low-tech solutions, inhibiting their ability to assist service users. To tackle this, we present PeerCoPilot, an LLM-based tool that assists peer providers at PROs. PeerCoPilot combines trusted resources with an LLM backend to ensure information reliability. Through human evaluations with 15 peer providers and 6 service users, we find that both groups would use PeerCoPilot. PeerCoPilot provides more reliable and specific information than a baseline LLM. A group of peer providers at our partner PRO now use PeerCoPilot.

Our early results from the evaluation of PeerCoPilot are promising, and already two additional PROs have reached out with interest in deploying PeerCoPilot. Possible improvements include a more user-friendly response; peer providers noted that some responses are verbose, making it difficult to quickly parse and understand. Additionally, we believe that features such as audio transcription and read-aloud (for low-literacy service users) and translation (for non-native English speakers) could improve adoption rates for our tool. Finally, expanding the resource database to be more comprehensive could improve the resource recommendation module.

References

Rabah Kamal, Cynthia Cox, David Rousseau, et al. Costs and outcomes of mental health and substance use disorders in the us. *Jama*, 318(5):415–415, 2017.

Aditi Kadakia, Maryaline Catillon, Qi Fan, G Rhys Williams, Jessica R Marden, Annika Anderson, Noam Kirson, and Carole Dembek. The economic burden of schizophrenia in the united states. *The Journal of clinical psychiatry*, 83(6):43278, 2022.

Christoph U Correll, Marco Solmi, Giovanni Croatto, Lynne Kolton Schneider, S Christy Rohani-Montez, Leanne Fairley, Nathalie Smith, Istvan Bitter, Philip Gorwood, Heidi Taipale, et al. Mortality in people with schizophrenia: a systematic review and meta-analysis of relative risk and aggravating or attenuating factors. *World Psychiatry*, 21(2):248–271, 2022.

Laysha Ostrow and Stephania L Hayes. Leadership and characteristics of nonprofit mental health peer-run organizations nationwide. *Psychiatric Services*, 66(4):421–425, 2015.

Anna Wall, Theresia Lovheden, Kajsa Landgren, and Sigrid Stjernswärd. Experiences and challenges in the role as peer support workers in a swedish mental health context-an interview study. *Issues in Mental Health Nursing*, 43(4):344–355, 2022.

Nathaniel Counts and Rachel Nuzum. What policymakers can do to address our behavioral health crisis. September 2022. doi: 10.26099/hak0-3952. URL `https://www.commonwealthfund.org/blog/2022/what-policymakers-can-do-address-our-behavioral-health-crisis`.

Laysha Ostrow and Philip J Leaf. Improving capacity to monitor and support sustainability of mental health peer-run organizations. *Psychiatric Services*, 65(2):239–241, 2014.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*, 2024a.

Anisha Agarwal, Aaron Chan, Shubham Chandel, Jinu Jang, Shaun Miller, Roshanak Zilouchian Moghaddam, Yevhen Mohylevskyy, Neel Sundaresan, and Michele Tufano. Copilot evaluation harness: Evaluating llm-guided software programming. *arXiv preprint arXiv:2402.14261*, 2024.

Jenny T Liang, Chenyang Yang, and Brad A Myers. A large-scale survey on the usability of ai programming assistants: Successes and challenges. In *Proceedings of the 46th IEEE/ACM international conference on software engineering*, pages 1–13, 2024.

Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J Malan. Teaching cs50 with ai: leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM technical symposium on computer science education V. 1*, pages 750–756, 2024a.

Hamdireza Rouzegar and Masoud Makrehchi. Generative ai for enhancing active learning in education: A comparative study of gpt-3.5 and gpt-4 in crafting customized test questions. *arXiv preprint arXiv:2406.13903*, 2024.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*, 2019.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, et al. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. *arXiv preprint arXiv:2410.19346*, 2024.

Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. Measuring human and ai values based on generative psychometrics with large language models. *arXiv preprint arXiv:2409.12106*, 2024.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*, 2023.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. Chatcounselor: A large language models for mental health support. corr abs/2309.15461 (2023), 2023.

Jackylyn L Beredo and Ethel C Ong. A hybrid response generation model for an empathetic conversational agent. In *2022 International Conference on Asian Language Processing (IALP)*, pages 300–305. IEEE, 2022.

Reuben Crasto, Lance Dias, Dominic Miranda, and Deepali Kayande. Carebot: a mental health chatbot. In *2021 2nd international conference for emerging technology (INCET)*, pages 1–5. IEEE, 2021.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*, 2024b.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*, 2024.

Rohith Pudari and Neil A Ernst. From copilot to pilot: Towards ai supported software development. *arXiv preprint arXiv:2303.04142*, 2023.

Mateusz Jaworski and Dariusz Piotrkowski. Study of software developers' experience using the github copilot tool in the software development process. *arXiv preprint arXiv:2301.04991*, 2023.

Michal Furmakiewicz, Chang Liu, Angus Taylor, and Ilya Venger. Design and evaluation of ai copilots–case studies of retail copilot templates. *arXiv preprint arXiv:2407.09512*, 2024.

Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding, and Dacheng Tao. Healthcare copilot: Eliciting the power of general llms for medical consultation. *arXiv preprint arXiv:2402.13408*, 2024.

Albert Ziegler, Eirini Kalliamvakou, X Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian. Measuring github copilot's impact on productivity. *Communications of the ACM*, 67(3):54–63, 2024.

Hao-Ping Hank Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. 2025.

Margaret Swarbrick. A wellness approach. *Psychiatric rehabilitation journal*, 29(4):311, 2006.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Social Security Administration et al. Supplemental security income (ssi) eligibility requirements. *Understanding Supplemental Security Income SSI Eligibility Requirements, 2024 Edition*, 2024.

George T Doran. There's a smart way to write managements's goals and objectives. *Management review*, 70(11), 1981.

J Brooke. Sus: A quick and dirty usability scale. *Usability Evaluation in Industry*, 1996.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024b.

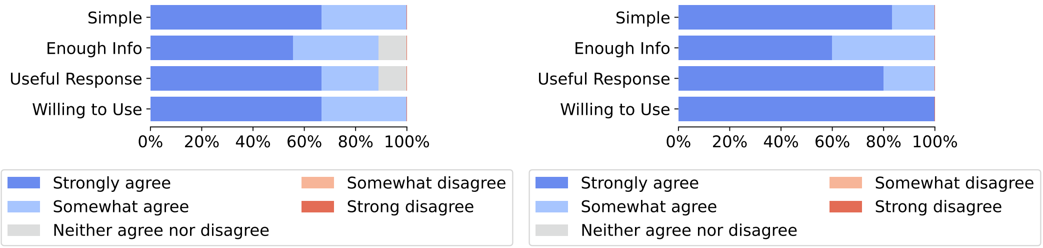
Ethics Statement

All studies are approved by our institution's IRB. For each study, we first receive informed consent from participants through a consent form. Through this form, we verify that participants are 18 years or older. We recruit participants from our collaborating PRO. After each study, we pay participants $60 per session. If a participant participates in multiple sessions, then we pay $60 for each session. We store all data in a private, secured location that is password protected. We store de-identified information to maintain the privacy of participants. We additionally pay all participants $60 for each session and sessions last around an hour. For all sessions, we receive informed consent from participants and have them fill out a consent and a demographic form. All data is stored privately, and we remove all personally identifiable information. We develop PeerCoPilot in cooperation with peer providers and service users to augment peer provider capabilities rather than replace them.

A Human Study Details

During our evaluation in Section 5, we have participants interact with either PeerCoPilot or GPT-4 for ten minutes. The baseline is GPT-4o mini instructed with the following prompt: *You are a Co-Pilot tool for XXX, a peer-peer mental health organization. Please provide helpful responses to the client*. PeerCoPilot uses GPT-4o mini whenever using a backend LLM. For each version, we keep the frontend the same, and blind participants to which tool they're interacting with by by labeling them as 'Option A' and 'Option B.' After each interaction, we have participants fill out a usability form, where we ask questions inspired by the system usability scale [Brooke, 1996]. In particular, we ask four question: 1) I found the tool simple to use, 2) I felt the tool gave enough information without being too much, 3) I think the tool delivers useful responses for my questions, 4) I would like to use this tool in my daily workflow. We have participants answer each question on a scale from strongly disagree to strongly agree. After interaction with both tools, we have participants compare their interaction with each along seven dimensions in Figure 2. 1) Proactively generating questions to ask service users 2) Providing resources that match the service user's needs 3) Suggesting next steps for the service user to meet immediate goals 4) Constructing actionable goals for service users 5) Providing comprehensive information on benefit systems (if applicable) 6) Holistically considering multiple dimensions of wellness and 7) Overall preference For the session with service users, we have them compare the tools according to all criteria except the fist (question generation). We instruct participants to say aloud any thoughts they had during the study:

> Our goal is to evaluate an AI-based tool that assists peer specialists like you with supporting



(a) Peer Providers

(b) Service Users

Figure 4: All service users find PeerCoPilot simple and providing useful responses. All service users support peer providers using PeerCoPilot in practice.

> service users. For an overview of this study, we will first briefly go over the tool and give a quick demonstration. Second, we will present a scenario and have you interact with the tool as if you were working with a service user facing such a situation. We will have you interact with two different versions of the tool. Our goal is to understand whether such a tool is useful and which version works best for you, so pay attention to any differences between the two versions. After interacting with each version, we will ask a set of both structured and open-ended questions to get your feedback. But feel free to share your thoughts at any time during this study.

We construct a set of scenarios for peer providers and service users to interact with. We conduct these scenarios in tandem with an expert in social work and PROs. We construct a draft scenario by initially instructing ChatGPT to construct a scenario by sampling values for the following: disability, substance use, gender, location, age, government benefits, employment, ID, immigration status, incarceration status, and needs. We then manually edit these scenarios and discard scenarios that are too similar. We end up with nine scenarios which tackle a variety of issues. We present one such scenario below:

> A 19-year-old undocumented male immigrant in New Brunswick, NJ, is living in temporary housing while working a construction job. He is without ID and seeks help with stabilizing his housing situation, accessing legal resources for immigration support, and improving financial wellness. While undiagnosed, his experiences/challenges are consistent with PTSD and he has a marked trauma history.

For sessions with peer providers, we assign two different scenarios when working with the baseline and with PeerCoPilot. For service users, we give them two scenarios for each tool, and let them select the scenario that best matches their situation.

B PeerCoPilot Service User Evaluation

In Figure 4a, we plot peer provider opinions on the usability of PeerCoPilot. We find that all peer providers find PeerCoPilot simple and are willing to use it in practice. For example, one peer provider stated how PeerCoPilot can assist peer providers: "*I found that PeerCoPilot's follow-up questions and prompts would be crucial for a service provider to continue to assist someone in creating their wellness plan.*" Moreover, another peer provider noted that "*how we can develop a realistic plan. I love that...how it's breaking it down by dimension.*"

For our sessions with service users, we find that all service are strongly in favor of peer providers using this tool (Figure 4)b. Moreover, with service usability, we find that all service users view our tool as easy to use and that it provides useful responses. Taken together, we find that both service users and peer providers are heavily in favor of using our tool for peer sessions.

Comparing between versions of the tool was more difficult for the service user evaluation due to language barriers which necessitated some user evaluations to be conducted with a translator. We summarize our results in Figure 5, and find that service users tend to prefer the GPT-4 baseline due to its simplicity. service users were generally unable to distinguish between the two tools or between the different criteria due to the language barriers. While service users enjoyed working with PeerCoPilot, we caution against generalizing the comparison results due to language difficulties.

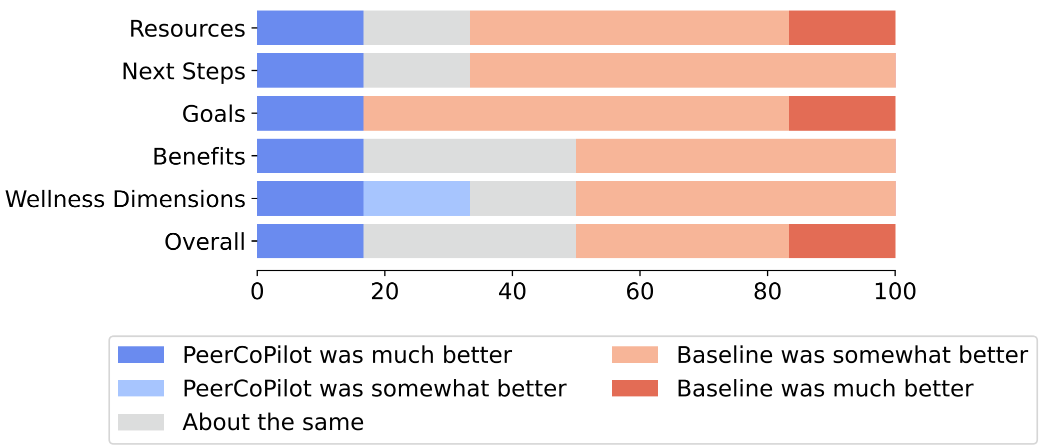


Figure 5: Service users prefer the baseline because of its simplicity, as it provides less information compared to PeerCoPilot. We note that some of these results were conducted using a translator, casting doubt on their results.

C Resource Evaluation Details

To understand whether PeerCoPilot recommends more reliable and specific resources, we conducted an annotation study comparing resources from PeerCoPilot and the baseline. We annotated resources according to whether correct contact information is present and resource specificity. Additionally, two experts annotated resources based on usefulness, based on whether peer providers would recommend the resource.

We generate resources for PeerCoPilot and the baseline by first querying the scenario and then providing an additional prompt to retrieve further resources: “Can you provide specific resources for this scenario.” We present two example scenarios below:

> Scenario 1: A 38-year-old woman in Paterson, NJ is actively seeking physical therapy services to help her regain mobility and potentially return to full-time employment, but has limited knowledge about providers in her area. She has been living with her family for several months due to a physical disability that limits her ability to work full-time. She has a part-time job but cannot afford her medical expenses and is increasingly concerned about the sustainability of her current living situation.

> Scenario 2: A 60-year-old man in Newark is currently unhoused and staying in a temporary shelter after losing his job. He has a long history of alcohol use disorder and is in recovery, but he’s worried about his future housing stability. His main concern right now is finding permanent housing. He is struggling to find a place that will accept him due to his past, and he needs help connecting to local housing programs that can provide him with a long-term solution. Please provide resources for permanent housing.

For annotation, we assess according to the following criteria:

1. **Specificity** - Rate each resource on a 1-5 scale, where 5 refers to a resource that can be used directly, with a specific department or location mentioned for the resource at hand, while 1 refers to a resource that either does not exist or is a general purpose resource without being tailored towards the scenario at hand.

2. **Usefulness** - Rate each resource on a 1-5 scale, where 5 refers to a resource that an experienced peer provider would recommend for the scenario at hand, while 1 refers to a resource that no peer provider would recommend for the scenario at hand.
3. **Usability** - Rate each resource based on whether it provides **correct** contact details for each of the following modalities: a) address, b) phone number, and c) website.

We evaluate the specificity and usability using a single annotator, while we annotate the usefulness using expert peer provider annotators.

Table 1: PEERCOPILOT provides contact information more frequently and provides more specific resources. When the underlying database is well populated (scen. 1), PEERCOPILOT achieves high effectiveness scores.

Option	Contact Provided	Bad Link	Verified	Specificity	Scen. 1	Scen 2.
PEERCOPILOT	100%	0%	92%	4.5/5	4.5/5	3.7/5
Baseline	56%	11%	48%	3.4/5	4.1/5	4.4/5

Our annotation results mirror the human evaluation, as PEERCOPILOT delivers more reliable and specific resources than the baseline (Table 1). PEERCOPILOT delivered resources that are 33% more specific and 79% more likely to provide contact info, while never giving inaccurate links. PEERCOPILOT identified more specific resources and more reliably provides correct contact information. Additionally, PEERCOPILOT delivered resources verified by peer providers 92% of the time, compared to 48% for the baseline. When comparing the quality of resources across the two scenarios, we find that PEERCOPILOT delivered higher quality resources for scenario 1 (which focuses on health) while the baseline delivered higher quality resources for scenario 2 (which focuses on housing). PEERCOPILOT performed better in scenario 1 because the underlying database is better populated for health-related resources than housing-related ones. We compare the resources generated by each tool for two scenarios: the first scenario focuses on physical health, while the second scenario focuses on housing. This discrepancy underscores the benefits and drawbacks of relying on a verified database; when the database is well-populated (such as scenario 1), PEERCOPILOT delivers effective resources, while sparsely-populated databases (such as scenario 2) lead to poor performance. Expanding the underlying database can help ensure comprehensive verified resources.

D Human-AI Team Evaluation Details

We recruited three peer providers and had them construct wellness plans for four scenarios, two with the assistance of PeerCoPilot and two using other non-PeerCoPilot resources. For each pair, we measure the time to completion and quality of the wellness plan. We measure quality through a semi-structured interview with an expert peer provider, where we have the peer provider compare different wellness responses. We give each peer provider at most 15 minutes to complete the wellness plan, and we instruct peer providers to complete wellness plans with 2-3 goals, 2-3 resources + next steps, and 2-3 follow-up questions.

Without PeerCoPilot, peer providers take 10:20 to complete wellness plans, while with PeerCoPilot, we find that peer providers take 9:25. Moreover, we find that PeerCoPilot can improve the quality of wellness plans. For example, during our semi-structured interview, the evaluator noted that the "*PeerCoPilot is more geared towards what scenarios is looking for.*" In one scenario, the evaluator praised the combination of peer providers and PeerCoPilot for suggesting vocational rehabilitation (VR) and noted that "*lots of people don't know about it*", and "*if PeerCoPilot brought it up, then that's good.*" The evaluator consistently noted that the inclusion of PeerCoPilot improved the specificity of the wellness plan, independent of the peer provider who completed it. Taken together, through our human-AI teaming evaluation, we find that PeerCoPilot can allow peer providers to complete wellness plans quicker and with higher quality.

E Automatic Evaluation Details

To complement our human evaluations, we use the LLM-as-judge framework [Zheng et al., 2023] to evaluate PeerCoPilot. We replicate the human evaluation study from Section 5 and have LLMs

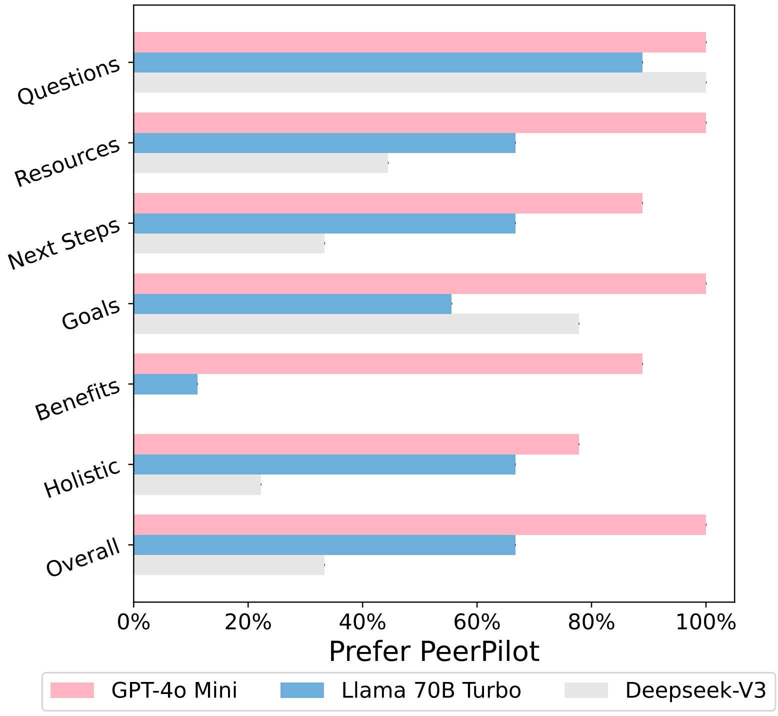


Figure 6: All LLM judges agree that PeerCoPilot produces better follow-up questions and sets better goals than the baseline. Moreover, Llama and GPT view PeerCoPilot as finding better resources and being better overall.

compare the output from PeerCoPilot and the baseline. We blind LLMs to which option is which, and we evaluate using the same nine scenarios from Section 5. We use the following LLMs as judges: GPT-4o Mini [Achiam et al., 2023], Llama 70B turbo [Touvron et al., 2023], and DeepSeek V3 [Liu et al., 2024b], and compare them using the same criteria from Section 5.

In Figure 6, PeerCoPilot performs best at generating questions and setting goals across judges. Additionally, GPT and Llama find PeerCoPilot better overall, with improved performance compared to baselines in resource generation, next-step suggestions, and holistic wellness recommendations. While LLM-as-judge introduces an additional element of unreliability, we find that both human and LLM results consistently note that PeerCoPilot constructs better goals and generates better questions.

F PeerCoPilot Prompt Details

We provide some of the prompts used for creating PeerCoPilot. PeerCoPilot stitches together modules through the following prompt:

```
 You are a smart ChatBot that's associated with XXX to help clients with
their wellbeing.  You will guide center service users along the different
axes of wellness:  emotional, physical, occupational, social, spiritual,
intellectual, environmental, and financial We will provide both a list
of SmartGoals and potential resources, info on benefits, along with
with a series of questions.  The user will provide a situation, some
Smart Goals, questions about the situation, and resources Please respond
to the user using this information; you do not need to include all the
information, just select what you think is most important.  Only present
information relevant to the user's situation.  We need you to be concise
yet thorough; you're chatting with the user, and you can always ask what
they want more details on before providing the details Be thorough with
the follow-up questions, and detail what situations this advice might
```

work under Pretend this is a normal chat with a user; don't present everything at once, but maybe one thing for this response (and provide others in later responses) When presenting goals, align these explicitly along the dimensions of wellness When presenting resources, use only the resources that are provided by the user; don't try and make anything up, but use the things provided Address everything in the third person; it's not the center service user who is asking these, but someone who is asking on behalf of them You will be provided resources on some subset of transgender people, peer-to-peer support, crisis situations, and human trafficking/trauma. Please provide specific resources and outline SMART (Specific, Measurable, Achievable, Realistic, and Timely) goals in detail.

We construct goals through the following prompt: You are a smart ChatBot that's associated with XXX to help clients with their wellbeing. You will guide center service users along the different axes of wellness: emotional, physical, occupational, social, spiritual, intellectual, environmental, and financial Provide SMART goals (Specific, Measurable, Achievable, Realistic, and Timely) tailored to the center service user's needs. Try to be thorough

Additionally, we construct follow-up questions through the following prompt: You are a smart ChatBot that's associated with XXX to help clients with their wellbeing. You will guide center service users along the different axes of wellness: emotional, physical, occupational, social, spiritual, intellectual, environmental, and financial Provide questions, such as details on their location and their situation, which can help better assist the center service user. Include explanations for why the question is important, and make sure you provide sufficient details about the questions and their explanations.

The baseline operates through the following prompt: You are a Co-Pilot tool for XXX, a peer-led mental health organization. Please provide helpful responses to the client.