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

Gao Mo[*,1], Naveen Raman[*,1], Megan Chai[1], Cindy Peng[1], Shannon Pagdon[2], Nev Jones[2], Hong Shen[1], Peggy Swarbrick[3], Fei Fang[1]

[1] Carnegie Mellon University, [2] University of Pittsburgh,
[3] Collaborative Support Programs of New Jersey

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 with their day-to-day tasks, 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 to support these goals. 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 5-10 peer providers at a leading PRO serving over 10,000 service users, and we are actively expanding PEERCOPILOT's use. [1]

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).

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).

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. 2024b; Agarwal et al. 2024; Liang, Yang, and Myers 2024). As a result, LLMs present a promising opportunity for PROs by potentially helping peer providers craft tailored wellness plans and synthesize location-specific 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, Yang, and Myers 2024), LLMs are rarely, if ever, used in PROs because many peer providers lack 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. We developed PEERCOPILOT in partnership with a leading PRO from the Northeast United States. After conversations with the PRO, we designed PEERCOPILOT to assist with common tasks faced by peer providers, such as wellness plan creation, goal construction, resource recommendation, and benefit navigation. In our discussions, peer providers also 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 more reliable and specific information than a baseline LLM. Our work is preliminarily deployed to a group of peer providers, who use PEERCOPILOT in their daily operations, and two other PROs have reached out as a result of our initial deployment.

[1]We will publicly release our code and tool after the review period

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. 2024b; Rouzegar and Makrehchi 2024; Rodriguez, Jafari, and Ormerod 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). Within behavioral health, most related are work in mental health on simulating patients (Wang et al. 2024a; Louie et al. 2024) and work on using AI to answer substance use questions (Giorgi 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.

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 System Design

We develop PEERCOPILOT as a web-based chatbot which peer providers can interact with and ask questions to. PEERCOPILOT combines an LLM backend with modules that rely on verified information sources to ensure reliability. We describe the overall backend before describing individual modules.

3.1 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).

3.2 Backend Modules

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 at our partner PRO. 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.

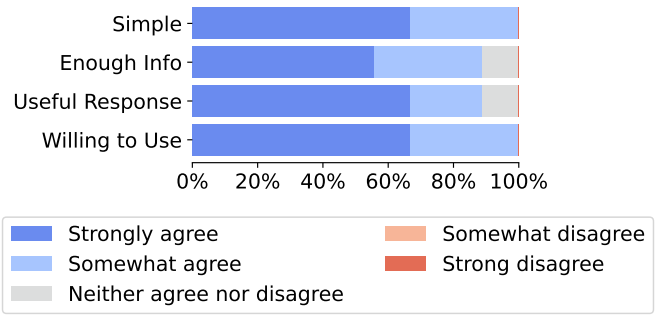

Figure 1: We surveyed peer providers on the usability of PEERCOPILOT. All peer providers found PEERCOPILOT simple and are willing to use it in practice.

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, in the benefit eligibility module, 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)), and can be updated as requirements change. This results in an assessment of whether a service user is likely to be eligible for each benefit.

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 partners at the 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 and additionally prompts with information on 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?"

4 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.

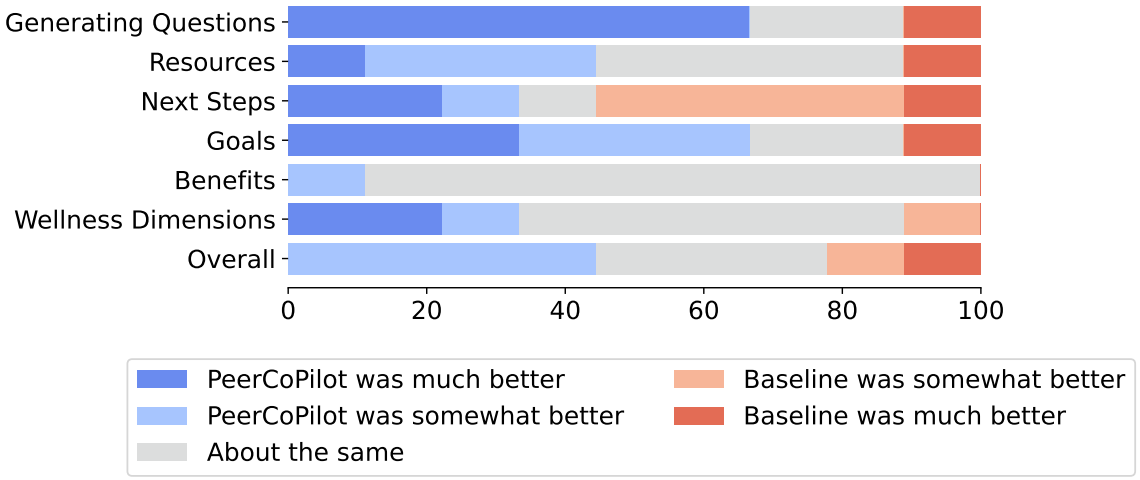

Figure 2: Peer providers find that PEERCOPILOT generates better questions, recommends better resources, and crafts better goals compared to a baseline. This is because the modules ensure information reliability and specificity.

4.1 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 diverse scenarios capturing different types of situations faced by service users in reality; 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 our comparisons against other LLMs through the LLM-as-judge framework.

Peer providers are willing to use PEERCOPILOT In Figure 1, we find all 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.*" Another peer provider said 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.*"

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 is...insane that it gave the birth certificate requirements.*" Conversely, for the baseline, one peer provider stated that they "*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?*"

4.2 Reliability and Specificity

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; details in Appendix C.

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 our partner PRO 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)

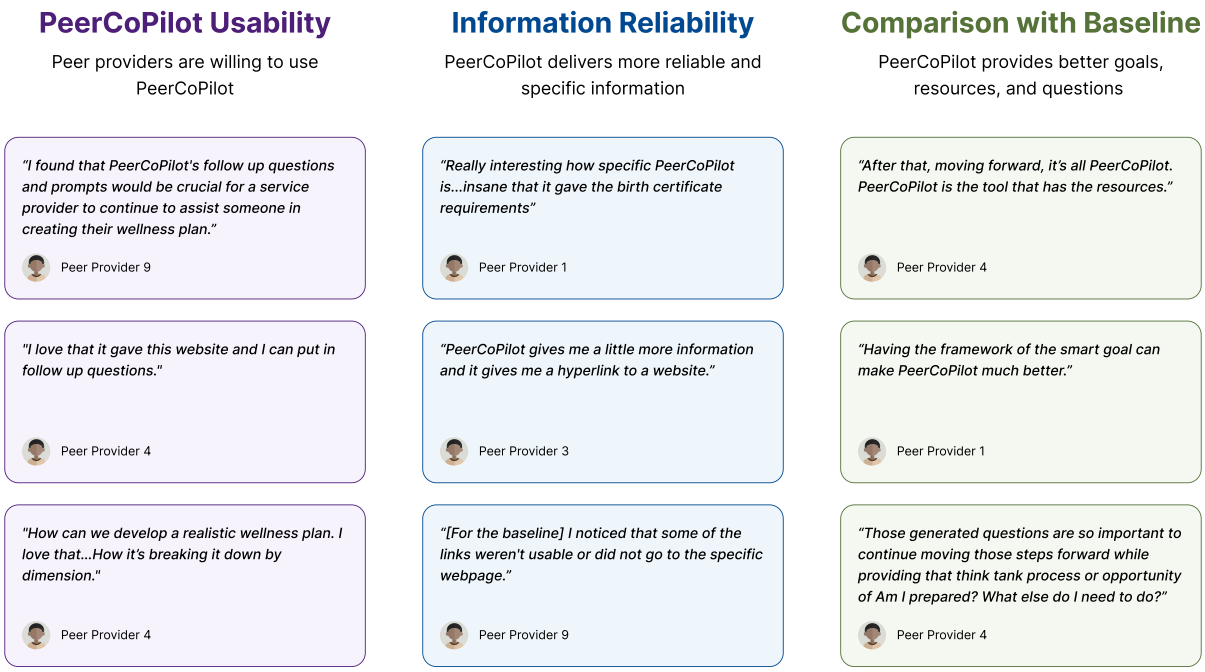

Figure 3: We outline three themes from our sessions with peer providers: 1) PEERCOPILOT provides useful information and peer providers are willing to use it, 2) PEERCOPILOT provides reliable and specific information, and 3) PEERCOPILOT provides better goals and questions than the baseline.

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

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 comprehensiveness.

4.3 Preliminary study on Peer provider-PEERCOPILOT Teaming

We conducted a pilot human-AI teaming study to assess PEERCOPILOT's ability to help peer providers create wellness plans. We assessed the completion time and quality for wellness plans completed with and without PEERCOPILOT. We show PEERCOPILOT reduced completion time by 10% while leading to wellness plans better tailored to service user situations (details in Appendix D). Our study demonstrates the positive impact PEERCOPILOT can have when peer providers use it in practice.

5 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. We find that PEERCOPILOT provides more reliable and specific information than a baseline LLM. A group of peer providers at our partner PRO use PEERCOPILOT regularly, and we working to expand this.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Agarwal, A.; Chan, A.; Chandel, S.; Jang, J.; Miller, S.; Moghaddam, R. Z.; Mohylevskyy, Y.; Sundaresan, N.; and Tufano, M. 2024. Copilot evaluation harness: Evaluating llm-guided software programming. *arXiv preprint arXiv:2402.14261*.

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

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

Correll, C. U.; Solmi, M.; Croatto, G.; Schneider, L. K.; Rohani-Montez, S. C.; Fairley, L.; Smith, N.; Bitter, I.; Gorwood, P.; Taipale, H.; et al. 2022. 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.

Counts, N.; and Nuzum, R. 2022. What Policymakers Can Do to Address Our Behavioral Health Crisis.

Crasto, R.; Dias, L.; Miranda, D.; and Kayande, D. 2021. CareBot: a mental health ChatBot. In *2021 2nd international conference for emerging technology (INCET)*, 1–5. IEEE.

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

Furmakiewicz, M.; Liu, C.; Taylor, A.; and Venger, I. 2024. Design and evaluation of AI copilots–case studies of retail copilot templates. *arXiv preprint arXiv:2407.09512*.

Giorgi, S.; Isman, K.; Liu, T.; Fried, Z.; Sedoc, J.; and Curtis, B. 2024. Evaluating generative AI responses to real-world drug-related questions. *Psychiatry research*, 339: 116058.

Jaworski, M.; and Piotrkowski, D. 2023. Study of software developers' experience using the Github Copilot Tool in the software development process. *arXiv preprint arXiv:2301.04991*.

Kadakia, A.; Catillon, M.; Fan, Q.; Williams, G. R.; Marden, J. R.; Anderson, A.; Kirson, N.; and Dembek, C. 2022. The economic burden of schizophrenia in the United States. *The Journal of clinical psychiatry*, 83(6): 43278.

Kamal, R.; Cox, C.; Rousseau, D.; et al. 2017. Costs and outcomes of mental health and substance use disorders in the US. *Jama*, 318(5): 415–415.

Lai, T.; Shi, Y.; Du, Z.; Wu, J.; Fu, K.; Dou, Y.; and Wang, Z. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Lee, H.-P. H.; Sarkar, A.; Tankelevitch, L.; Drosos, I.; Rintel, S.; Banks, R.; and Wilson, N. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Liang, J. T.; Yang, C.; and Myers, B. A. 2024. 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*, 1–13.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, J. M.; Li, D.; Cao, H.; Ren, T.; Liao, Z.; and Wu, J. 2023. ChatCounselor: A Large Language Models for Mental Health Support. CoRR abs/2309.15461 (2023).

Liu, R.; Zenke, C.; Liu, C.; Holmes, A.; Thornton, P.; and Malan, D. J. 2024b. 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*, 750–756.

Louie, R.; Nandi, A.; Fang, W.; Chang, C.; Brunskill, E.; and Yang, D. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*.

Mou, X.; Liang, J.; Lin, J.; Zhang, X.; Liu, X.; Yang, S.; Ye, R.; Chen, L.; Kuang, H.; Huang, X.; et al. 2024. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. *arXiv preprint arXiv:2410.19346*.

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

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

Pudari, R.; and Ernst, N. A. 2023. From copilot to pilot: Towards AI supported software development. *arXiv preprint arXiv:2303.04142*.

Ren, Z.; Zhan, Y.; Yu, B.; Ding, L.; and Tao, D. 2024. Healthcare copilot: Eliciting the power of general llms for medical consultation. *arXiv preprint arXiv:2402.13408*.

Rodriguez, P. U.; Jafari, A.; and Ormerod, C. M. 2019. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*.

Rouzegar, H.; and Makrehchi, M. 2024. 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*.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867.

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

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wall, A.; Lovheden, T.; Landgren, K.; and Stjernswärd, S. 2022. 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.

Wang, R.; Milani, S.; Chiu, J. C.; Zhi, J.; Eack, S. M.; Labrum, T.; Murphy, S. M.; Jones, N.; Hardy, K.; Shen, H.; et al. 2024a. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

Wang, R. E.; Ribeiro, A. T.; Robinson, C. D.; Loeb, S.; and Demszky, D. 2024b. Tutor CoPilot: A human-AI approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*.

Ye, H.; Xie, Y.; Ren, Y.; Fang, H.; Zhang, X.; and Song, G. 2024. Measuring human and ai values based on generative psychometrics with large language models. *arXiv preprint arXiv:2409.12106*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Ziegler, A.; Kalliamvakou, E.; Li, X. A.; Rice, A.; Rifkin, D.; Simister, S.; Sittampalam, G.; and Aftandilian, E. 2024. Measuring GitHub Copilot's impact on productivity. *Communications of the ACM*, 67(3): 54–63.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291.

Ethics Statement

We conduct all user studies under 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 partnering 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 4, 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 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 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 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:

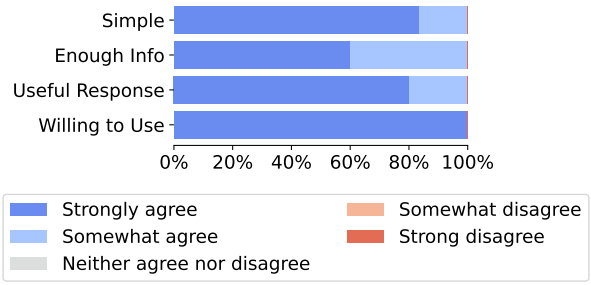

Figure 4: All service users find PeerCoPilot simple and providing useful responses. All service users support peer providers using PeerCoPilot in practice.

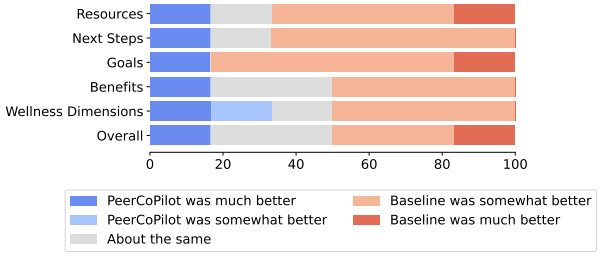

Figure 5: Service users prefer the baseline because of its simplicity, as it provides less information compared to PeerCoPilot. We note that some of these results were conducted using a translator, casting doubt on their results.

> 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

For our sessions with service users, we find that all service are strongly in favor of peer providers using this tool (Figure 4). Moreover, with service users, we find that all service users view our tool as easy to use and that it provides useful responses. Taken with Figure 1, 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.

C Resource Evaluation Details

We provide further details on the evaluation in Section 4.2. 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.

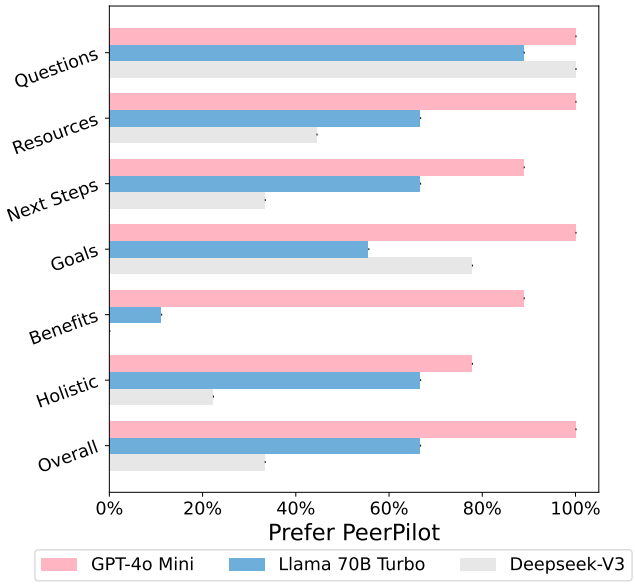

Figure 6: All LLM judges agree that PeerCoPilot produces better follow-up questions and sets better goals than the baseline. Moreover, Llama and GPT view PeerCoPilot as finding better resources and being better overall.

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 4 and have LLMs 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 4. 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. 2024a), and compare them using the same criteria from Section 4.

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 helps 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 helps 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 that helps 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 a peer-led mental health organization. Please provide helpful responses to the client.