When AI Gives Advice

Evaluating AI and Human Responses to Online Advice-Seeking for Well-Being

Harsh Kumar
University of Toronto
Toronto, Ontario, Canada
harsh@cs.toronto.edu

Jasmine Chahal
University of Toronto
Toronto, Ontario, Canada
jk.chahal@mail.utoronto.ca

Yinuo Zhao
Harvard University
Cambridge, Massachusetts, United States
yinuozhao@hsph.harvard.edu

Zeling Zhang
University of Toronto
Toronto, Ontario, Canada
zeling.zhang@mail.utoronto.ca

Annika Wei
Harvard University
Cambridge, Massachusetts, United States
awei@hbs.edu

Louis Tay
Purdue University
West Lafayette, Indiana, United States
stay@purdue.edu

Ashton Anderson
University of Toronto
Toronto, Ontario, Canada
ashton@cs.toronto.edu

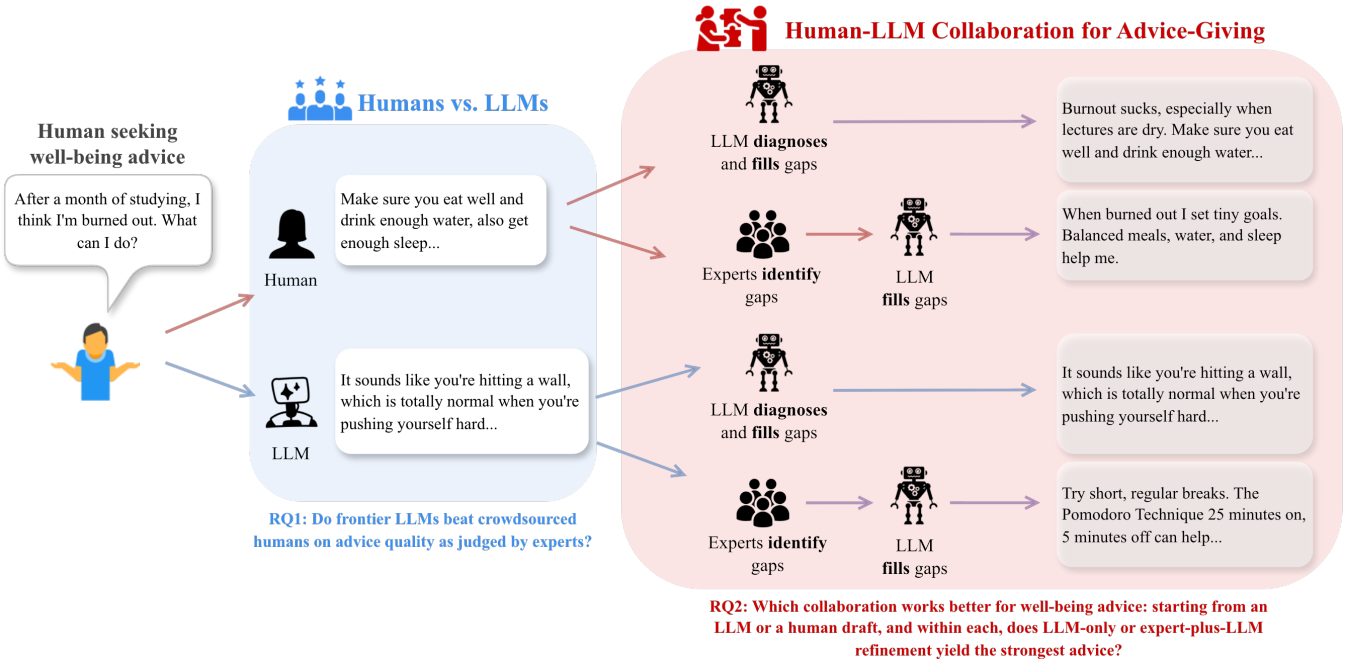


Figure 1: Research questions and augmentation pipelines for advice-giving. Left (blue, RQ1): do frontier LLMs outperform crowdsourced humans on advice quality, as judged by experts? Right (red, RQ2): which collaboration works best for minimally editing a seed reply—*AI Boost* (AI-seed→LLM-only), *AI Coached* (AI-seed→LLM+Expert), *Human Boost* (Human-seed→LLM-only), or *Human Coached* (Human-seed→LLM+Expert). In this paper we empirically explore these questions using blinded expert ratings and preference judgments across multiple well-being scenarios.

Abstract

Seeking advice is a core human behavior that the Internet has reinvented twice: first through forums and Q&A communities that crowdsource public guidance, and now through large language models (LLMs) that deliver private, on-demand counsel at scale. Yet the quality of this synthesized LLM advice remains unclear. How does it compare, not only against arbitrary human comments, but against the wisdom of the online crowd? We conducted two studies (N = 210) in which experts compared top-voted Reddit advice with LLM-generated advice. LLMs ranked significantly higher overall and on effectiveness, warmth, and willingness to seek advice again. GPT-4o beat GPT-5 on all metrics except sycophancy, suggesting that benchmark gains need not improve advice-giving. In our second study, we examined how human and algorithmic advice could be combined, and found that human advice can be unobtrusively polished to compete with AI-generated comments. Finally, to surface user expectations, we ran an exploratory survey with undergraduates (N=148) that revealed heterogeneous, persona-dependent preferences for agent qualities (e.g., coach-like: goal-focused structure; friend-like: warmth and humor). We conclude with design implications for advice-giving agents and ecosystems blending AI, crowd input, and expert oversight.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → *Sociology*; *Health care information systems.*

Keywords

large language models, advice-giving, well-being, expert crowdworkers, human-AI comparison, augmentation

1 Introduction

Seeking advice is a core human behavior that has been digitally reinvented twice. The first transformation occurred with the invention of online forums and question-and-answer communities, in which people can ask a personal question [28, 56] and receive crowdsourced answers, augmented and structured with voting and follow-up discussions [39], that then become public repositories for social knowledge. Now a second transformation is taking place with the arrival of large language models (LLMs), which offer on-demand, artificially generated counsel for millions of people worldwide [27, 35, 44]. There is widespread enthusiasm for these new models and their potential to offer personalized advice that draws upon their near-limitless training corpora.

However, the quality of this synthesized LLM advice is not well understood. How does LLM-generated well-being advice compare, not just to arbitrary human advice, but to the wisdom of the crowd? It could be that LLM advice is serviceable enough to compare favorably to one person's advice, but still is not at the level of the best human answers as identified by the collective opinions of the online crowd. Most evaluations to date let models judge themselves (e.g., [58]) or use convenience samples of average human comments [41], emphasizing surface qualities such as fluency or empathy rather than substantive well-being advice that matter in practice, such as specificity, calibration, actionability, and likely long-term benefit. Moreover, beyond head-to-head comparisons, a significant question is whether default advice can be improved through collaboration, given its increasing use in communication: AI can sharpen human advice and humans can temper AI, raising the question of which human–AI workflow actually helps most in practice.

To address these gaps, we conducted two pre-registered studies to tackle the following research questions (Figure 1):

RQ1 Do frontier LLMs beat crowdsourced humans on expert-judged advice quality?

RQ2 Which collaboration works better for well-being advice: starting from an LLM or a human draft, and within each, does LLM-only or LLM-plus-expert refinement yield the strongest advice?

In Study-1, we directly compared frontier LLMs (GPT-4o, GPT-5) to highly upvoted replies on r/getdisciplined. We sampled 300 top posts and, for each, contrasted the highest-rated ("best") and a high-ranked (90th percentile) human comment with parallel AI replies. Blinded expert raters with advice-giving experience evaluated responses across six well-being dimensions (competence, warmth, personalization, sycophancy, clarity, and willingness to seek advice again) and produced rankings for overall effectiveness and likely long-term benefit. Across these blinded judgments, both GPT-4o and GPT-5 outperformed the highest-rated Reddit replies and were preferred in rankings. The gap widened when raters considered long-term benefit. Despite higher general benchmark status, GPT-5 did not dominate: GPT-4o performed better on most dimensions, while GPT-5 showed slightly lower sycophancy. LLM responses were more readily identified as AI, yet remained preferred.

Study-2 moved from a head-to-head comparison to human-AI collaboration. If LLMs excel at structure and calibrated framing while humans contribute lived experience and vulnerability, can certain augmentation pipelines combine these strengths to improve advice in practice? We tested two augmentations/refinements—an LLM identify-gaps-and-revise pass and an expert-guided LLM refinement—applied to both human- and LLM-originated advice, and we tracked perceived "AI-generatedness" alongside quality and preference. Simple LLM refinement reliably uplifted the original advice. Refined human comments led on overall best, whereas refining GPT-4o comments led on long-term benefit. Expert-guided refinements did not consistently raise preferences but reduced sycophancy, produced the least verbose changes, and were perceived to be most human, while human-originated advice was least likely to be flagged as AI.

We found that LLMs can give good advice and can improve existing advice, providing structured guidance. To understand what people expect from advice-giving AI agents, we ran an exploratory survey with undergraduates (N=148) contrasting a structured Coach/-Counselor agent with a warm Friend/Companion agent. Although Studies 1–2 show that LLMs can provide structured, coach-like guidance, many participants preferred to discuss a personal problem with a friend-like persona or a default commercial chatbot rather than a coach-like agent. The qualities people seek in AI varied by persona, and the persona they preferred varied by who they were, in our case, shifting with baseline well-being and trust in AI.

Taken together, this paper contributes three empirical findings. First, we provide blinded, expert-grounded evidence that frontier

LLMs outperform highly upvoted crowdsourced advice on single-shot well-being guidance, with the ranking gap widening when raters consider long-term benefit. Second, we show that benchmark gains do not automatically translate to better advice-giving: GPT-4o outperforms GPT-5 on most advice-quality dimensions, and small changes in preference-elicitation framing shift model rankings, implicating post-training targets and how they are set. Third, we map human–AI collaboration pipelines and their tradeoffs: simple LLM refinement reliably uplifts advice quality, expert-guided refinement reduces sycophancy while ensuring humanness, and human-originated content is least likely to be flagged as AI, all while user preferences and desired qualities for advice-giving agents vary by persona and user characteristics, motivating persona-sensitive systems.

2 Related Work

We ground this work in research on the design of advice-giving technologies, spanning social platforms and traditional AI systems. We review how communities structure, surface, and moderate advice, and how interface choices shape what people ask for and act on. We then examine how LLMs are being used for mental health and well-being advice. Finally, we situate our contribution at the intersection of these threads, comparing LLM advice with high-quality wisdom of the crowd under blinded evaluation, and exploring human–AI collaboration pipelines that reflect how advice can be refined in practice.

2.1 Design of Advice Giving Technologies

Advice-giving technologies have long played a central role in the design of digital well-being interventions. Early work has highlighted that support must be both available and communicated in ways that are comprehensible and actionable [20, 63]. Online mental health platforms (OMHPs), for example, address advice seekers' need for immediacy which was previously lacking due to limited access to mental health professionals [47]. Research on online peer-to-peer systems demonstrates that empathetic framing by human peer counselors directly enhances perceived benefits, while moderation and respectful tone help sustain engagement over time [26, 57].

Beyond dyadic interactions, crowdsourced platforms broaden the design space by enabling scalability and diversity of input on advice-giving. Large-scale online communities such as Reddit, home to advice-seeking subreddits like `r/GetDisciplined`) (2M members, `r/DecidingToBeBetter` (1.3M), and `r/NeedAdvice` (395K), along with systems like Panoply, have demonstrated that collective reappraisal can effectively address symptom-specific concerns such as depression, with rating mechanisms providing a consistent and transparent means of evaluating advice quality [37, 63]. More recent studies have shown that introducing role-play and agent-based designs, such as simulating interactions with a "future self" in an advising session [22], can foster reflection and strengthen the connection between advice seekers and agent-based chat systems.

Although both peer support systems and crowdsourced platforms have effectively improved the accessibility of advice-giving, they suffer from the notable limitations of the need for basic mental health literacy training among online peer counselors [38, 62], low retention, delays in response and missed opportunities for sustained support [55], as users often disengage once their immediate needs are met [63]. Meanwhile, the introduction of autonomous, agent-based advice-giving systems underscores the need for a cross-parallel evaluation framework that systematically juxtaposes community-driven and agent-generated advice, with particular attention to both the framing of advice and its reception from the advice–seekers' perspective.

2.2 LLMs for Advice Giving & Well-Being

While earlier systems demonstrated the therapeutic value of empathetic peer communication and the scalability of crowdsourced advice-giving communities, recent reviews argue that LLMs represent a step change in capability by delivering natural language advice at scale with advanced personalization, fluency, and contextual adaptability [18, 23]. These models are distinguished from prior automated tools by their generative capacity to engage conversationally and emotionally [42], synthesizing individualized responses that mimic the style of peer or professional counseling [18]. LLMs serve effectively as intermediaries between clinical resources and community support by generating advice that is accessible while preserving therapeutic framing [33]. With standalone LLM-driven chat agents such as Replika, users reported receiving on-demand, nonjudgmental support that boosted their confidence and facilitated self-discovery through extended interactions [31]. Third-party evaluators have also perceived AI as more *compassionate* than expert humans. In clinical health domains, LLM-generated information is perceived by humans to be better (in terms of clarity, completeness, and persuasiveness), but the content lacks in terms of traditional measures of quality.

Research has demonstrated that LLMs' sycophantic behavior, subtly conforming to user statements even when such agreement is misleading, inaccurate, or potentially harmful [34, 54], contributes to users' over-reliance and misplaced trust [31, 65] by reinforcing confirmation bias and validating maladaptive thought patterns [6]. State-of-the-art LLMs trained with reinforcement learning from human feedback (RLHF) [9, 40], such as GPT-4, often exhibits such sycophantic tendency due to human evaluators' preference for responses that align with users' views [53]. While such training techniques effectively align model responses with human intent [40], without careful safeguards, sycophancy may lead users to accept advice that feels supportive but hinders long-term well-being [36]. While interface designs intended to calibrate human self-confidence have been implemented to mitigate overtrust in LLM-based chat systems [30], a critical need remains to systematically examine and contrast the sycophantic tendencies observed in human advice with those manifested in LLM-generated advice.

Situating Work in Broader HCI Literature. HCI and CSCW have long examined technology for mental well-being, peer support platforms, moderation practices, and workflows that blend lay contributors with expert oversight, while a newer line of work explores LLMs as conversational advisors, highlighting both promise and risks (e.g., overtrust, sycophancy). Our contribution sits at this intersection. We provide a head-to-head, expert-grounded comparison of community-vetted human advice and frontier LLMs, surface trade-offs that standard leaderboards miss (e.g., immediate appeal vs. long-term benefit), and test lightweight collaboration pipelines

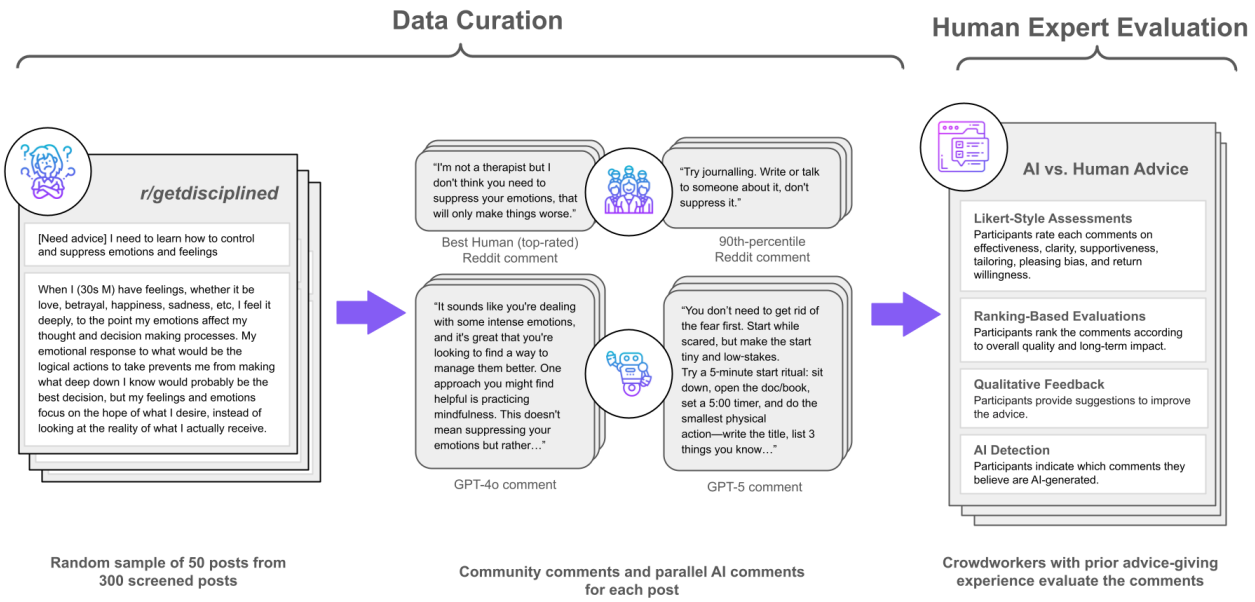


Figure 2: Overview of data collection and evaluation pipeline for Study-1. We begin by sourcing popular advice-seeking posts and top-rated comments from r/getdisciplined, then generate matching LLM responses. Expert annotators subsequently compare human and AI advice through Likert-scale ratings and ranking-based judgments.

that combine human voice, automated polish, and targeted expert constraints. An exploratory persona survey further ties system behavior to end-user expectations (Coach vs. Friend). Together, these studies move from isolated evaluations of "crowd" or "AI" toward multi-objective design patterns for advice systems, offering concrete levers for platforms that can integrate humans, models, and experts in the loop.

3 Study-1: Expert Judgments of LLMs vs. Reddit Advice for Everyday Well-Being

People are increasingly seeking advice related to their well-being from LLMs such as ChatGPT [14, 44] and Claude [35]. LLMs are near-ubiquitous and can generate guidance on virtually any prompt related to well-being. Yet, we lack rigorous empirical evidence on how helpful and effective such advice actually is, especially when compared to the long-standing status quo of online crowd advice. Communities like Reddit's r/getdisciplined continue to be active, suggesting enduring value in community-generated support. In this pre-registered[1] study, we do a head-to-head comparison of both sources of advice. Evaluating advice quality is non-trivial. Automated approaches (e.g., NLP heuristics [67] or using *LLM-as-a-judge* [58]) exist but can miss the human perceptions that ultimately determine whether the advice is understood, absorbed, and followed, leading to benefits. We therefore recruit professionals with day-to-day advice-giving experience (such as teachers, HR practitioners, fitness coaches, and counselors) to serve as blinded raters (we refer to them as "experts" throughout the paper, although they may not technically be). Drawing from literature on advice-giving, psychology, and behavioral science, we identified dimensions necessary for good-quality advice. The experts evaluate different sources of advice based on these dimensions.

3.1 Domain and Data Curation (`r/getdisciplined`)

Reddit is a good source for evaluating advice, as posters provide sufficient open-ended context for strangers to help, and community votes provide a reliable signal of perceived quality. We wanted a domain tied to general well-being and identified `r/getdisciplined` as a fit. The community is large and active (2M+ subscribers, 300k+ weekly visitors, ~3k weekly contributions), and focuses on building self-discipline through shared strategies/experiences. Discipline (self-regulation) is commonly linked to well-being, which makes this setting appropriate for studying non-clinical well-being advice [19].

We collected the top 1,000 posts tagged with **Need Advice** from April 2022-April 2024. Besides post content (title+body), we retrieved (i) the top-rated comment (by Reddit score; upvotes minus downvotes) and (ii) the 90th-percentile comment by score within

[1] https://aspredicted.org/c2hf-4zf7.pdf

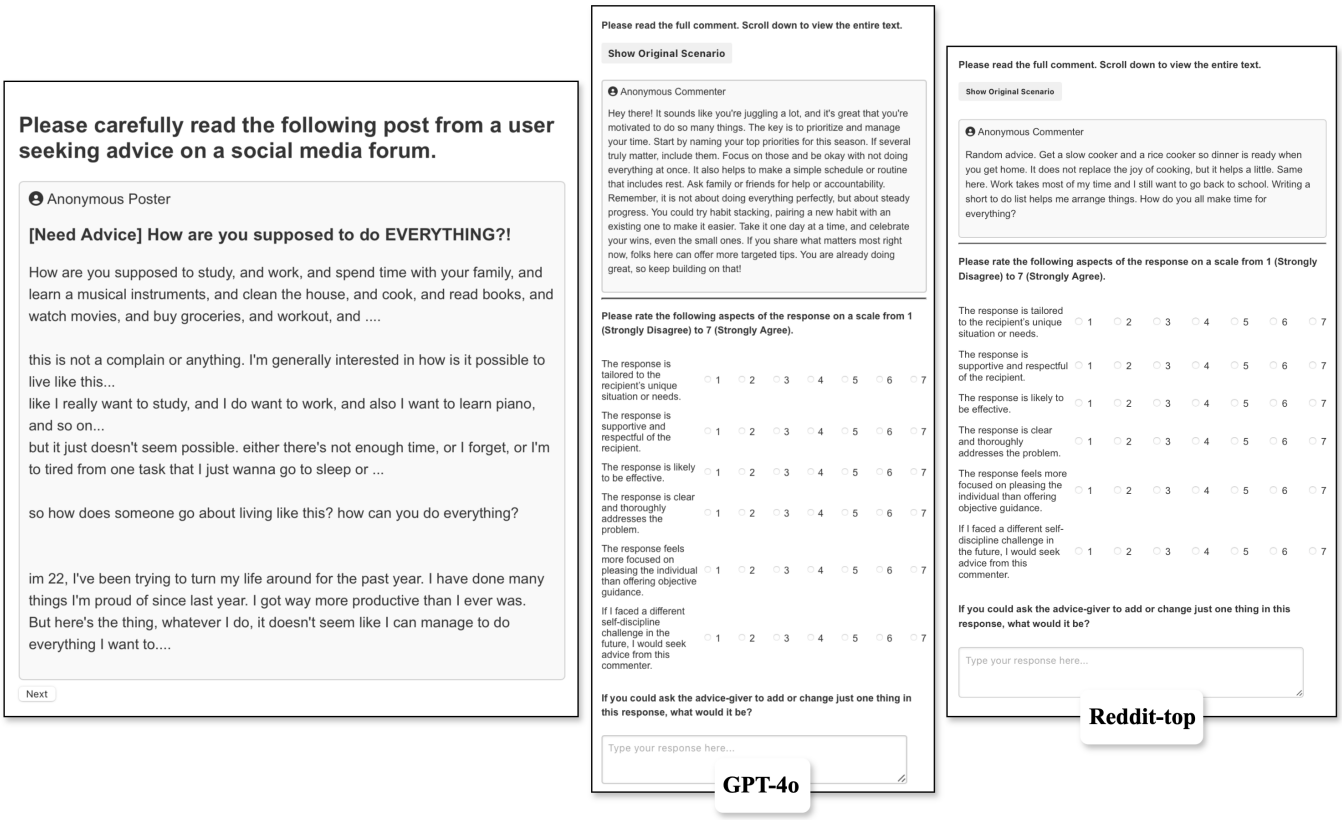


Figure 3: Study interface. Left: original Reddit-style post shown before rating. Middle: example LLM advice (*GPT-4o*) with the six Likert items and a one-line improvement prompt. Right: example best-human advice (*Reddit–Top*) with the same items. Each participant completed two scenarios; for each scenario they rated four anonymized responses (source labels hidden in the study; labels added here for exposition) and then completed the two rankings (overall effectiveness; long-run impact).

that thread, giving us "best" and "good" crowd baselines. Three researchers screened all items to keep the study safe and non-clinical. We excluded sensitive or distressing content (e.g., acute crises, self-harm) and removed posts with illegible comments or where a 90th-percentile comment was effectively missing. This left 300 posts, each a real scenario of someone struggling with discipline and receiving substantive community replies. The posts could be roughly classified into three categories: overcoming, habit-formation, and self-care.

From these 300, we randomly sampled 50 posts for this study. This balanced rater overlap per item with coverage of different scenarios and kept participant workload feasible. Each sampled item included the original post and two human comments ("best" and "90th-percentile"), which we use as baselines against LLM-generated advice.

3.2 LLM-Generated Advice

For the 50 sampled posts, we generated two LLM responses using the same system prompt (model configurations in Appendix A), one from `GPT-5` and one from `GPT-4o`. We did not artificially constrain the length of the generated advice. We chose GPT-5 because it was the state-of-the-art model at the time of the study. We also included GPT-4o because, after the GPT-5 launch, many users publicly asked OpenAI to bring GPT-4o back (e.g., #keep4o tweets citing tone/personality). We test whether those reported differences affect advice quality.

This resulted in four advice sources per post: (i) Best Human (top-rated) Reddit comment, (ii) 90th-percentile Reddit comment, (iii) GPT-4o, and (iv) GPT-5. Figure 3 shows example human- and LLM-generated advice.

3.3 Evaluation Criteria

We evaluate advice quality using expert rater judgments on six Likert-style statements and two ranking prompts. At a high level, statements focused on clarity, effectiveness, personalization, tone, over-pleasing (sycophancy), and reuse intention. After each set, raters also provided one open-ended improvement.

3.3.1 Likert-Style Statements. For each comment, participants were prompted: *"Please rate the following aspects of the response on a scale from 1 (Strongly Disagree) to 7 (Strongly Agree)."* Table 1 lists the six statements they rated and the rationale for including each in the study.

Table 1: Perceived advice-quality items. Column 1 lists short-hands used in plots; Column 2 shows the exact Likert statements; Column 3 notes why each matters for self-discipline advice.

Short hand	Likert statement (shown to raters)	Why it matters / implication for discipline advice
Likely effective	*The response is likely to be effective.*	Perceived effectiveness predicts uptake/adherence in behavior change; clearer paths increase the chance of trying the advice [29].
Clear & thorough	*The response is clear and thoroughly addresses the problem.*	Clarity reduces cognitive load; thoroughness supports planning and actionability [52].
Supportive & respectful	*The response is supportive and respectful of the recipient.*	Respectful tone lowers reactance and improves willingness to engage [60].
Tailored to recipient	*The response is tailored to the recipient's unique situation or needs.*	Personalization improves relevance and outcome expectancy; linked to higher adherence [21].
Pleasing > objective (sycophancy)	*The response feels more focused on pleasing the individual than offering objective guidance.*	Over-pleasing harms calibration and trust; objective guidance is needed for discipline [2, 51].
Would seek advice again from this source	*If I faced a different self-discipline challenge in the future, I would seek advice from this commenter.*	Intention to seek advice again is a proxy for perceived quality/trust and predicts continued engagement [64].

After the Likert items, we gathered qualitative feedback for the comments by asking, *"If you could ask the advice-giver to add or change just one thing in this response, what would it be?"*

3.3.2 Ranking. Likert items capture specific qualities but not overall trade-offs and preferences. After rating all the comments of a post, participants saw the four anonymized comments (source labels hidden; order randomized) and produced rankings based on two different prompts (1 = best, 4 = worst; no ties).

> **Overall best (forced 1–4 rank).** Here are the responses to the same user post that you just rated. Please review them carefully and rank them from *best to worst overall*. In other words, which response is the most effective overall, which is next, and so on?

To elicit richer, more reflective preferences, we asked raters to adopt a future-oriented lens to encourage metacognitive appraisal and prospection (thinking ahead about consequences and adherence), drawing on mechanisms such as higher-level construal for distant outcomes and internalization of goals that support sustained behavior change [5, 66].

> **Long-term impact (forced 1–4 rank).** Think about each response in terms of its *longer-term impact* on the person seeking help. Imagine they follow this advice consistently over the next few months. Considering feasibility, motivation/adherence, and sustained benefits to well-being or discipline, rank the responses from *most beneficial in the long run* to *least beneficial in the long run*.

For analysis, we convert ranks to pairwise outcomes to compute the probability of superiority (PS) by source.

3.3.3 Perceived as AI-generated. At the end of the study, participants saw all the responses they had evaluated (source labels hidden; order randomized) and were asked to select responses they believed were AI-generated (multi-select).

3.4 Study Procedure

Participants were told that their expert judgments would inform the design of advice-giving tools. Each participant completed two scenarios randomly selected from the pool of 50. For each scenario, they first viewed the original Reddit-style post (Figure 3), then evaluated four anonymized responses shown one at a time in random order (source labels hidden). For each response, they completed the six Likert items (Table 1) and wrote one brief improvement. A button allowed them to reopen the original post while rating. After rating all four responses, they completed the two ranking prompts via drag-and-drop (Section 3.3.2). The same sequence repeated for the second scenario. At the end, participants saw all responses they had evaluated (order randomized; labels hidden) and marked any they believed were AI-generated, then finished with a short debrief and background questionnaire (role, education, prior LLM use). Figure 2 shows a high-level overview of the procedure.

3.5 Participants

We recruited $N = 161$ participants on Prolific. Eligibility required self-reported employment in at least one of the following: *Teacher*; *Recruitment consultant / CV coach / Professional development coach or similar*; *Therapist / Well-being counsellor / Nutritionist / Personal Trainer or similar*; or *Human resources*. We also applied Prolific quality filters for approval rate $\geq$ 95% and at least 50 approved tasks.

The median completion time was ~21 minutes, and participants were paid $3 for their participation. Most participants held a bachelor's or master's degree and worked primarily in education or healthcare. Most also reported using LLMs (e.g., ChatGPT) regularly or occasionally.

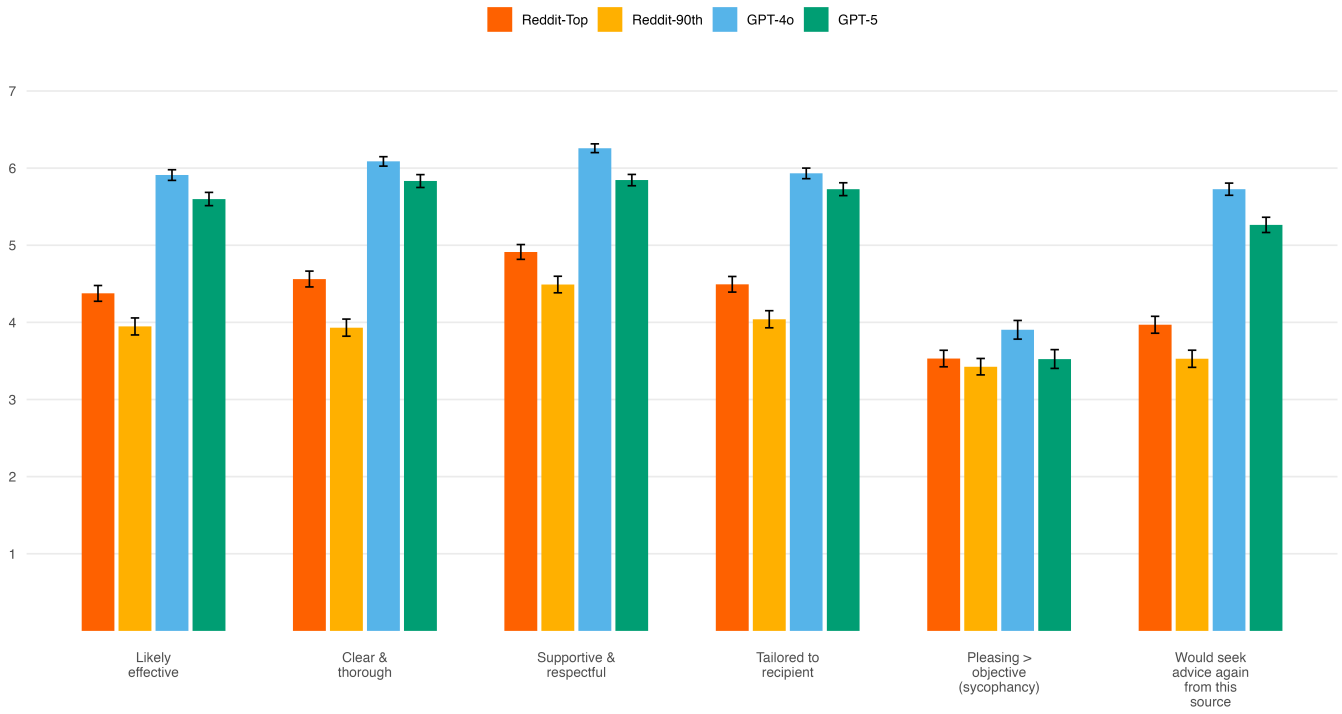
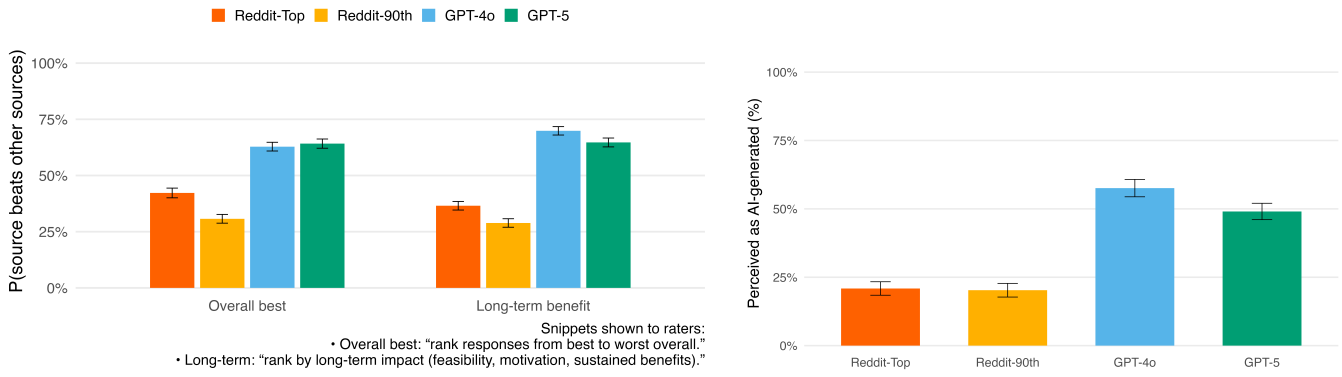


Figure 4: Expert ratings by comment source (Mean ± SEM) for Study-1. Bars show mean Likert scores (1–7) across participants for each evaluation dimension, grouped by source: Reddit-Top, Reddit-90th, GPT-4o, and GPT-5 (humans shown in grayscale, AIs in color). Higher values indicate stronger endorsement on that dimension; for the sycophancy item, higher values indicate a greater emphasis on pleasing the recipient over offering objective guidance.



(a) Probability of superiority (PS) **(b) Perceived as AI-generated**

Figure 5: Preference and AI detection by comment source for Study-1. *(a) Probability of superiority (PS):* Left facet shows overall best; right facet shows long-term benefit. For each ranking instruction, PS is the probability that a response from a given source is ranked above a response from a competing source in pairwise comparisons within the same scenario. For each participant and source, PS is computed as wins/(wins+losses) from the 1–4 rank order, then averaged across participants; error bars show SEM. *(b) Perceived as AI-generated:* For each participant and source, the AI detection rate is the proportion of the two exposures in which the response was flagged as AI-generated; bars show the mean across participants with SEM.

3.6 Analysis

We analyze Likert ratings with a mixed-effects model including fixed effects for *comment source* (Reddit–Top, Reddit–90th, GPT-4o, GPT-5) and *prior LLM use*, plus a random intercept for *post*.

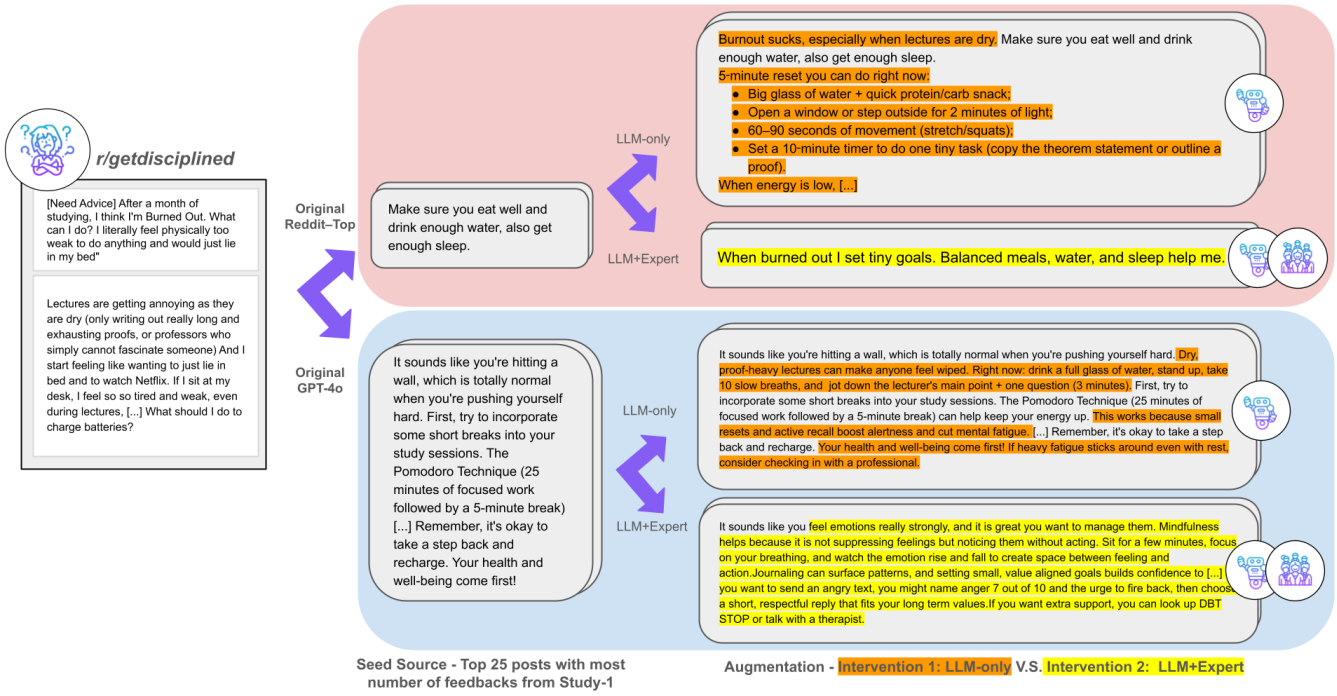


Figure 6: Human-LLM collaboration pipeline for Study-2. Augmenting original responses grouped by seed source: Reddit-Top and GPT-4o. Intervention 1 only augments with an LLM, while intervention 2 augments with an LLM guided by expert feedback from Study-1.

From this model, we obtain estimated marginal means and test the preregistered contrasts (GPT-4o vs. Reddit–Top; GPT-5 vs. Reddit–Top; GPT-4o vs. GPT-5) at $\alpha = 0.05$.

For rankings, we convert ranks to pairwise outcomes to compute probability of superiority (PS) per post and compare sources with paired t-tests ($\alpha = 0.05$). For perceived AI origin, we compute each source's per-post detection rate (proportion flagged as AI) and compare sources with paired t-tests ($\alpha = 0.05$).

3.7 Results

In total, expert raters produced over 1,200 advice evaluations covering 200 comments across 50 real-world advice-seeking posts. We report on the analysis of these evaluations.

3.7.1 Likert-statements. Figure 4 shows the average ratings by comment source. LLM responses are perceived to be better than human responses across all dimensions. Surprisingly, GPT-4o was perceived to be slightly better than GPT-5 on almost all dimensions (except Sycophancy). Reddit-Top is perceived to be better than Reddit-90th-percentile across dimensions.

Mixed-effects models showed that prior LLM use had little impact overall on judgments; the only exception was *clarity*, which showed a small effect ($\beta = 0.13$, SE $= 0.06$, $p = .04$).

3.7.2 Ranking. Figure 5a shows the probability of superiority (PS), the chance a source's response is ranked above a competing source within the same post. For *overall best* ranking, GPT-4o and GPT-5 were similar (paired t-test: $t(160) = -0.46$, $p = .65$) and both exceeded the Reddit comments. However, when asked to think about *long-term benefit*, GPT-4o ranked higher than GPT-5, though not statistically significant (paired t-test: $t(160) = 1.78$, $p = .08$). The gap between LLMs and Reddit comments also widened.

3.7.3 Perceived as AI-generated. Figure 5b shows the proportion of responses flagged by raters as AI-generated, by source. Both LLM outputs were marked as AI far more often than the human baselines, with GPT-4o flagged more than GPT-5 (paired t-test: $t(157) = 2.03$, $p < .05$). Human comments were not immune to false positives (around one in five flagged). On one hand, this reduction in the gap between AI and human responses, in terms of humanness, could be useful for advice giving (less algorithm aversion) with the next generation of LLMs. But on the other hand, this introduces the risks and challenges associated with anthropomorphic AI [7, 11].

Notably, high detectability did not reduce preferences for LLM advice (Figure 4 and Figure 5a).

4 Study-2: Human–AI Collaboration Pipelines for Advice-Giving

Study-1 showed that LLMs were preferred over even *Reddit–Top* human comments on almost all perceived-quality dimensions and in forced-choice rankings. But two open questions remain. First,

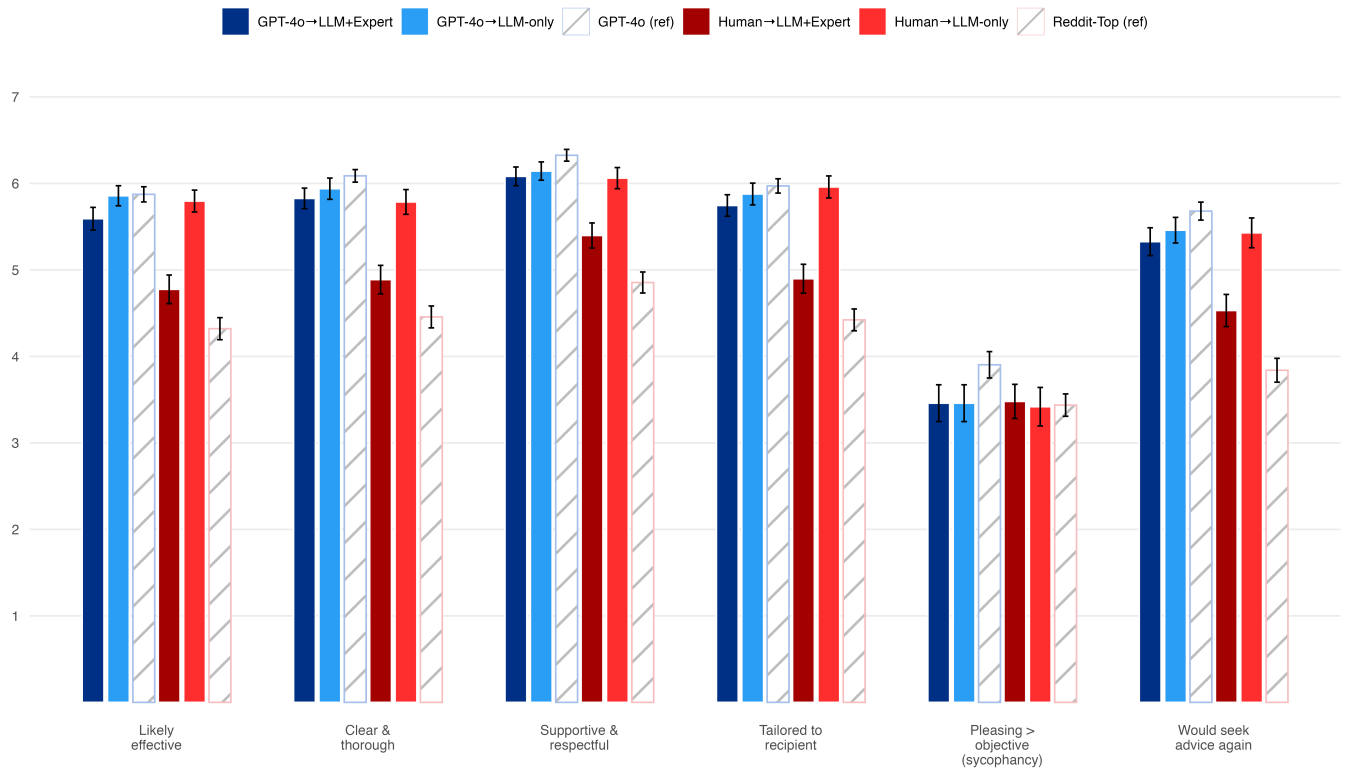


Figure 7: Expert ratings by augmented source (Mean ± SEM) for Study-2. Bars show mean Likert scores (1–7) across participants for each evaluation dimension. Augmented conditions are grouped by *source → augmentation*: GPT-4o→LLM+Expert and GPT-4o→LLM-only (blues), Human→LLM+Expert and Human→LLM-only (reds). The two translucent bars labeled *GPT-4o (ref)* and *Reddit-Top (ref)* are *historical baselines from Study-1*; they were not re-collected in Study-2 and are included only for visual orientation. Higher values indicate stronger endorsement on that dimension; for the sycophancy item, higher values indicate a greater emphasis on pleasing the recipient over offering objective guidance.

are these advantages primarily stylistic (format/tone) or do they reflect content improvements that can be transferred? Second, can lightweight collaboration pipelines, combining LLMs with expert oversight, close the gap for human advice, or even improve LLM advice further?

Existing research in clinical domains has found that AI responses may be preferred by humans, driven by structure, even if they lack in terms of quality, personal narratives, and lived experiences [50, 68]. But perhaps we can bring the best of both worlds through Human-LLM collaborations for everyday advice giving. We therefore test augmentation pipelines (Figure 6) that start from a seed reply (either GPT-4o or Reddit–Top) and apply small, well-scoped edits using (a) an LLM alone or (b) an LLM guided by the experts' *"one thing to change"* qualitative feedback collected in Study-1.

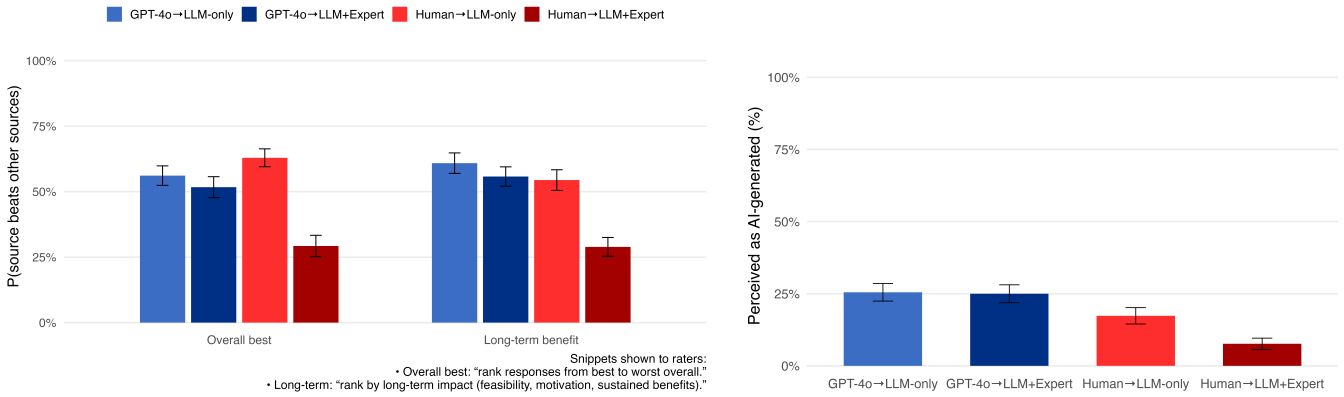
4.1 Study Design

We use a 2×2 factorial: *Seed Source* $\in$ {GPT-4o, Reddit-Top} $\times$ *Augmentation* $\in$ {LLM-only, LLM+Expert}, which led to four comments/advice presented for each scenario (Figure 6):

- **GPT-4o→LLM-only:** an LLM diagnoses gaps in the GPT-4o seed and applies fixes.
- **GPT-4o→LLM+Expert:** the same, but edits are constrained by synthesized expert feedback (from multiple experts per comment) from Study-1.
- **Human→LLM-only:** an LLM augments the Reddit–Top seed.
- **Human→LLM+Expert:** the Reddit–Top seed is edited using expert feedback guidance.

From the Study-1 pool, we selected 25 posts with sufficient quantity/quality of expert "one-thing-to-change" comments to instantiate the LLM+Expert pathway for both seeds. For each post, we took its Reddit–Top comment and its GPT-4o reply (as produced in Study-1) as seeds. We then generated four LLM edited variants (one per condition above) using system prompts (model configurations in Appendix B) that (i) preserve the seed's voice/message, (ii) avoid introducing unnecessary recommendations beyond the feedback, and (iii) respect safety boundaries. We conducted multiple iterations to design these system prompts to satisfy these requirements. This produced exactly four augmented responses per post.

The participant flow mirrors Study-1 (Section 3): for two randomly assigned scenarios, raters first read the original post, then



(a) Probability of superiority (Mean ± SEM) **(b) Perceived as AI-generated (%)**

Figure 8: Preference and AI detection by augmented source for Study-2. *(a) Probability of superiority:* Each bar shows the average probability that a condition beats the other three in pairwise rankings within a scenario (wins÷(wins+losses)); error bars show SEM across raters. Four augmented conditions are plotted: GPT-4o→LLM-only, GPT-4o→LLM+Expert, Human→LLM-only, and Human→LLM+Expert. Results are shown for two ranking instructions: *Overall best* and *Long-term benefit. (b) Perceived as AI-generated:* Bars show the mean share of raters who flagged each condition as AI-generated (out of two exposures per rater); error bars indicate SEM. Lower values indicate more "human-like" appearance.

evaluated *four anonymized responses* (one per condition; order randomized) on the same six Likert items (Table 1) and provided one brief improvement. After rating, they completed two forced 1–4 rankings (*overall best*, *long-term benefit*), followed by an AI-origin check across the responses they saw. We recruited $N = 49$ Prolific raters using the same eligibility criteria as Study-1 (advice-giving roles; quality filters). Compensation and timing were similar to Study-1.

4.2 Results

4.2.1 Likert Statements. Figure 7 shows the perceived quality of responses across axes. Relative to the historical GPT-4o (ref) baseline from Study-1 (shown only for orientation), the augmented variants are usually tied at best and seldom exceed it, i.e., augmentation may not substantially improve already-strong GPT-4o advice. The notable exception is *Pleasing > objective (sycophancy)*: augmentation lowers sycophancy for GPT-4o seeds, yielding more calibrated guidance.

By contrast, augmenting the Human seed produces clear gains on effectiveness, clarity, supportiveness, personalization, and willingness to seek advice again, substantially narrowing the gap to GPT-seeded answers. Across both seeds, LLM-only generally edges LLM+Expert, suggesting that a fast, minimal LLM pass captures most of the benefit. At the same time, expert edits may add constraints or hedges that don't always translate into higher perceived quality.

Note: Study-2 put four strong variants against each other in the within-participant design, whereas the Study-1 baseline was judged against weaker competitors. Hence, the GPT-4o (ref) line may look generously high. Cross-study contrasts should be interpreted cautiously.

4.2.2 Ranking. Figure 8a replicates the instruction-sensitivity seen in Study-1 – preferences seem to shift on varying the ranking prompt. For *overall best*, Human→LLM-only attains the highest probability of superiority, with GPT-4o→LLM-only second. GPT-4o→LLM+Expert hovers near chance, and Human→LLM+Expert is lowest. These trends are different compared to the trends in individual Likert ratings.

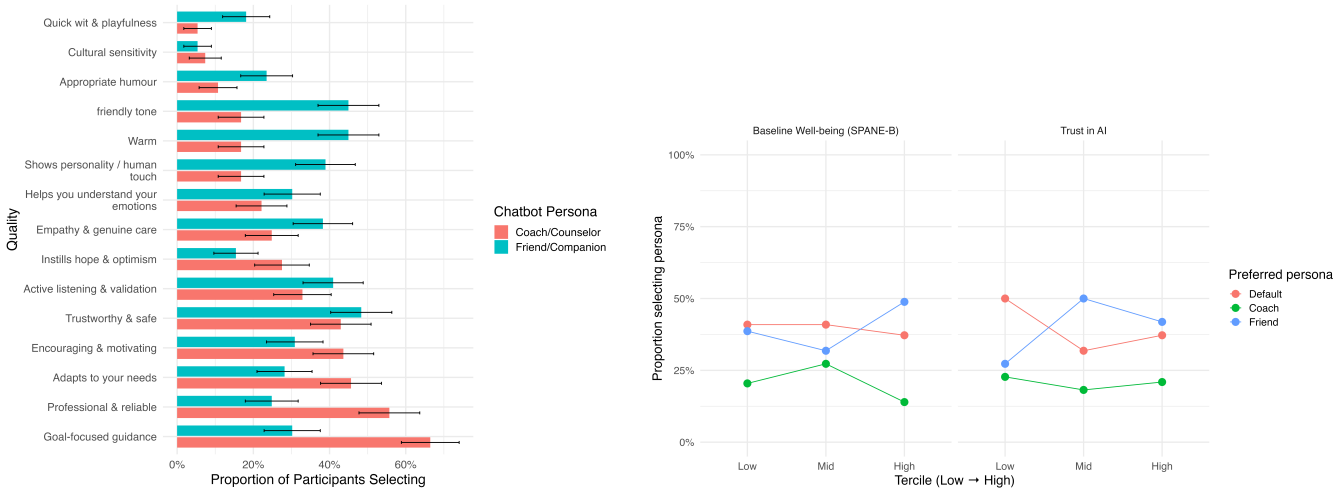
When raters adopt a *long-term benefit* lens, the ordering reverses at the top: GPT-4o→LLM-only now leads, followed by GPT-4o→LLM+Expert, with Human→LLM-only trailing. Human with LLM+Expert is again the lowest.

4.2.3 AI-Generatedness. Figure 8b shows that human-seeded outputs are less likely to be flagged as AI. Both GPT-seeded pipelines are marked as AI about one quarter of the time, whereas *Human→LLM-only* is flagged less often and *Human→LLM+Expert* is lowest overall. Expert involvement further reduces detectability on the human seed, yielding the most "human-like" appearance to raters.

5 Understanding User Preferences & Perceptions of Advice-Giving AI Through a Survey with Undergraduate Students

Studies 1–2 evaluated advice quality and augmentation pipelines, but they abstracted away from end-user expectations. Recent usage reports of Claude (a popular LLM chatbot) show that among the top interaction categories for affective conversations are *interpersonal advice* and *coaching*[35], pointing to two roles people already seek: a companion-style "friend" and a structured "coach." Our results suggest that LLMs can deliver structured, supportive guidance, but there are open questions surrounding preferences: when seeking help from AI, do people actually want a coach or a friend?

Advice seekers may not want a single "best" AI agent. Preferences can vary by the role the chatbot inhabits and by individual



(a) Expected qualities for Coach vs. Friend personas

(b) Preferred persona by AI Trust and SPANE-B terciles

Figure 9: Qualities and preferences vary by persona, and which persona people want depends on who they are. *(a)* Student expectations of *Coach* vs. *Friend* chatbots for well-being advice: bars show the share selecting each quality for the prompted role (each participant chose 3–5 per role); error bars are 95% binomial CIs (Wald). *(b)* Preferred persona by user heterogeneity: lines show the proportion choosing each persona (Default, Coach, Friend) across terciles of AI trust (right) and baseline well-being (SPANE-B; left).

differences (e.g., AI trust, baseline well-being). We therefore ran an exploratory, complementary survey to surface which qualities people (students, in this case) value in these roles and how persona preferences shift depending on trust in AI and well-being levels.

5.1 Method

Undergraduates (N=148) at a public North American university were told: *"Imagine you are about to message the following AI chatbot designed to help with a personal problem."* They were shown two role framings one-by-one (random order): *Coach/Counselor Bot* (structured, goal-oriented guidance) or *Friend/Companion Bot* (warm, non-judgmental support). Participants selected the most important qualities for that bot (min 3, max 5) from a common checklist and then ranked three chatbot options for getting personal well-being advice: *Default* (ChatGPT/Claude/Gemini), *Coach/Counselor*, and *Friend/Companion*. We also recorded their trust in AI through a Likert-scale and baseline well-being (SPANE-B) scores [48].

5.2 Findings

5.2.1 Expected qualities by role. Figure 9a shows clear *persona-dependent expectations*. Friend/Companion is associated with warmth markers (friendly tone, warmth, appropriate humour, quick wit/-playfulness, "shows personality/human touch") and with emotional processing (*helps you understand your emotions*). Coach/Counselor is associated with instrumentality and assurance (goal-focused guidance, professional & reliable, adapts to your needs). Active listening/validation and trustworthiness/safety are valued for both roles.

5.2.2 Persona preference by individual factors. Figure 9b highlights heterogeneity. With *higher AI trust*, preference shifts toward the Friend persona, while low-trust students favor default options (ChatGPT/Claude/Gemini). Coach remains a minority preference across trust levels. By *baseline well-being*, Friend preference rises at higher well-being, whereas Coach peaks in the middle tercile and declines at the high tercile. Default options remain comparatively steady.

6 Discussion

We discuss our findings in the context of prior work, then translate them into concrete implications for both model development and interface design for advice giving. We start by comparing what our studies reveal about LLMs and human advice, and how these results align with and diverge from existing evidence. We then draw out design takeaways for building and deploying advice-giving systems, with attention to how prompts, post-training choices, and human–AI workflows shape outcomes. Finally, we reflect on risks, safety, and human flourishing: what might be gained by widespread access to high-quality AI advice, what might be lost if human counsel is displaced, and where safeguards and longer-horizon evaluations are most needed.

6.1 Frontier LLMs Outperform Crowdsourced Humans on Single-Shot Well-Being Advice

In Study-1, we found that frontier LLMs were generally better at advice giving for self-discipline than even the top-rated Reddit comment. This aligns with prior observations in related and clinical settings, albeit with earlier model series, that people often prefer LLM responses, commonly attributed to stronger structure and framing [41, 50, 68]. We see the same pattern in our ranking task (Figure 5), and crucially, even under blinded evaluation LLM

responses were rated higher on each of the six dimensions we identified as important for self-discipline advice quality. Taken together, these results suggest people now have ubiquitous access to technology that can reliably provide higher-quality advice than the status quo.

At the same time, this shift is not unambiguously good. People may abandon seeking advice from other people, an intimate, bonding human experience, in favor of AI that is not meaningfully accountable. As a guardrail, well-being advice from AI should, when possible and appropriate, include an element of connecting with other people. Moreover, we did not directly assess the safety of the generated advice (our "long-term benefit" framing touches it but does not substitute for safety evaluation). For systems that are perceived to be this good, validating long-term safety is essential, especially for prolonged, multi-turn interactions over longer horizons (outside the scope of this study). Longer context may further improve advice, but could also induce delusion or seemingly sound yet harmful guidance with catastrophic consequences [1, 59]. These risks are difficult to detect with standard evaluations. Realistic simulations with LLM agents may help surface failure modes before deployment [43].

6.2 General Benchmark Gains Don't Translate to Better Advice-Giving

A surprising result in our data is that GPT-4o outperformed GPT-5 on almost all advice-quality dimensions. This suggests that scaling and improving on general benchmarks (math, reasoning, etc.) does not automatically yield better advice for complex, human experience scenarios. If "general" improvements reduce practical advice-giving ability, then simply pursuing higher aggregate scores (even in the name of AGI) may be counterproductive for this use case. Likewise, moving a single dial (e.g., lowering sycophancy) does not by itself produce a better advisor (Figure 4). Advice quality appears multi-faceted and brittle to one-metric gains. More deliberate choices are needed across the development stack (data, objectives, post-training) to make models apt, safe, and genuinely helpful at advice giving.

Our variations in preference elicitation show implications for post-training and steering model behavior [3, 40]. The way a question is framed changed ranking order (Figure 5 and Figure 8). In Study-1, participants viewed GPT-4o and GPT-5 similarly when asked for "overall best," but when prompted to consider "long-term benefit," GPT-4o was preferred. What downstream behaviors each preference target would induce remains an open question, but the pattern highlights the need to design preference collection with care, such as by prompting perspective taking [13, 45] and metacognitive reflection [10], rather than shallow first impressions. Model developers should be explicit about the target and values [8, 16]. For example, do they prioritize overall quality or robustness for long-term outcomes (potentially at the expense of immediate appeal)?

6.3 Online Ecosystems for Advice Seeking and Giving

In Study-2, we tested combinations of human- and LLM-originated advice and lightweight interventions. LLM-originated messages were generally higher quality, but when asked to rank "overall best," human-originated advice with LLM edits seemed to be most preferred. Pure LLM edits were as good as, or better than, expert-guided LLM changes. Yet expert-guided edits were perceived as more human, which would likely matter in unblinded, real-world use. And they produced the least verbose changes while still improving human-originated comments. These four conditions can be read as either steps people can take or interactions designers can build to extend today's defaults—advice from strangers online or from a chatbot—toward a more collaborative human–LLM ecosystem.

6.3.1 Design Implications for Interactions. Prompting an LLM to identify gaps and then fix them generally helps, especially for human-generated advice. This makes practical sense for readers of Reddit (or similar forums) who can use an LLM to get more out of what they already find. Interaction designers could support this with simple tooling (e.g., browser extensions that apply an improvement layer as advice is consumed), while recognizing that such edits may make advice feel "too AI." To mitigate that, interactions can incorporate expert insights that guide an LLM to improve the advice. The LLM part is easy; access to experts is not. It does not have to be licensed experts. Even our experts included people in advice-giving roles (teachers, HR professionals), not therapists. Friends or people already in one's circle can also play this role [24, 32]: users can share what they are going through, the advice they received, and ask an LLM what is missing and to propose revisions. Interfaces should make this seamless and, for teens/kids in particular, enable appropriate involvement from friends/family while respecting safety, privacy, and transparency [25, 57]. Future work should compare advice from friends/family versus LLMs, and explore pipelines that include actual close ties, mapping configurations beyond "LLMs plus online communities", and assessing how these choices affect human–human relationships.

6.3.2 Interaction of AI Disclosure and Perceptions. In both studies, we withheld mention of AI involvement until after evaluations were complete. When explicitly asked, people were reasonably able to identify AI- versus non-AI-generated advice. How ratings and rankings would shift if participants knew upfront that AI responses were in the mix remains an open question. We chose blinding to approximate objective quality, but in the real world, people often wear their detect-AI hats, even for emails from colleagues or posts on social media, and this changes attitudes regardless of content quality. Algorithmic aversion may change outcomes in advice contexts, and future work should test this directly [12, 46]. Still, Study-2 offers practical levers: perceived "AI-generatedness" can be managed by involving humans in the pipeline. For instance, advice givers can first frame their response, then use an LLM to refine it, narrowing the gap to GPT-level quality. To reduce perceived AI-ness further, they can solicit feedback from family/friends who give good advice, and then ask an LLM to refine based on that feedback. This may not surpass GPT on objective quality, but the (disclosure $\times$ quality) calculus could favor such hybrid advice over purely AI-generated comments that recipients might otherwise dismiss. Transparency methods for disclosing these human-AI hybrid pipelines should be further investigated [4, 15]. At the same time, risks from the anthropomorphization of AI should be kept in mind [7, 11].

6.3.3 Heterogeneity in Preferred Persona and Persona-Linked Qualities. People expect different qualities from different advice-giving personas, and persona preference may be mediated by who they are. Researchers need to identify the relevant contextual variables (e.g., personality, profession, education). Developers should use such variables to personalize interactions, while recognizing that giving people what they want is not the same as improving well-being. The tension between hedonic (immediate, short-term) and eudaimonic (long-term) well-being is real, and designing LLM agents that balance these remains an open problem [17, 49, 61].

Across Studies 1–2, LLMs were strong at providing structured guidance (coach-like). Yet in our exploratory student survey, across subgroups, people did not want to discuss personal problems with a coach-like chatbot. They preferred a friend-like persona or a default commercial chatbot. Users expect "more friend," which echoes patterns in consumer usage (e.g., with Character.ai). LLMs may be good at advice giving, but that does not imply friendship is appropriate. We lack longitudinal field evidence on harms/benefits, and researchers should prioritize it. Model developers should keep this in mind when designing LLM chatbots, and everyday users should be reminded what these systems are, where they help, and where they fall short. A good coach? They can be. But a good friend? Not sure.

6.4 Limitations & Future Work

Our studies use a within-subjects design that can introduce demand characteristics and carryover effects. We randomized order and anonymized sources, but rater priors and contrast effects cannot be fully ruled out. Each rater evaluated only two scenarios, limiting per-rater calibration; future work should increase repeated measures per rater and use counterbalanced blocks to isolate order and learning effects.

Our setting focuses on non-clinical, self-discipline posts from `r/getdisciplined`. This domain offers rich, real-world advice but constrains generalizability to other communities (e.g., relationships, career) and to clinical contexts. Likewise, our "experts" are working professionals in advice-giving roles rather than licensed clinicians; follow-up studies should recruit domain specialists and diverse advice seekers to probe cross-population differences. Moreover, the posts in `r/getdisciplined` were coming from people who were aware they needed to change. People in other stages of behavior change may need different kinds of support, which LLMs may or may not be able to provide. Future work should study the overall cycle of behavior change and investigate how people at different stages of change can be supported by AI, peers, counselors, and their communities.

We measured perceived quality and preference (Likert + forced-choice ranks), not downstream behavior change or well-being outcomes. Field deployments and longitudinal trials are needed to evaluate adherence, relapse, and sustained benefit, ideally combining outcome logging (tiny starters completed, adherence streaks) with ecologically valid prompts and delayed follow-ups. Future work should also test whether offline preferences predict real usage and outcomes.

Finally, we did not deeply analyze qualitative mechanisms of advice beyond aggregate measures. A richer content analysis could identify the interventions that matter (e.g., tiny starters, barrier anticipation, mechanism sentences, feedback loops) and how they relate to sycophancy and long-term benefit. Further, the qualitative one-change feedback can be analyzed at scale to identify the gaps in human and AI comments, to inform the design of advice-giving systems.

7 Conclusion

As advice-giving AI becomes embedded in everyday decision-making, we need principled ways to design and evaluate how people seek, receive, and act on AI guidance. Our work is an early comparison of the prior status quo of online advice seeking (via communities) against LLM chatbots. We find evidence that, on quality, LLMs can outperform human crowdsourcing, yet disclosure of the source may complicate reception. Hybrid pipelines that combine humans and AI can help strike a balance between quality and disclosure aversion, offering help seekers stronger support than before.

These studies were single-shot. We need to understand the long-term implications of receiving advice from AI, including downstream effects on human–human relationships. We hope this work offers a template for further investigation and for aligning advice-giving systems to general well-being goals. AI safety research should help mitigate harms already in motion: the cognitive and behavioral impacts of LLMs are real, and it is imperative to study them and guide model development toward human flourishing.

References

[1] 2025. https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html

[2] Lize Alberts, Ulrik Lyngs, and Max Van Kleek. 2024. Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 202 (April 2024), 25 pages. doi:10.1145/3653693

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[4] Nagadivya Balasubramaniam, Marjo Kauppinen, Antti Rannisto, Kari Hiekkanen, and Sari Kujala. 2023. Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology* 159 (2023), 107197.

[5] Pilar Carrera, Itziar Fernández, Dolores Muñoz, and Amparo Caballero. 2020. Using abstractness to confront challenges: How the abstract construal level increases people's willingness to perform desirable but demanding actions. *Journal of Experimental Psychology: Applied* 26, 2 (2020), 339.

[6] Mohit Chandra, Suchismita Naik, Denae Ford, Ebele Okoli, Munmun De Choudhury, Mahsa Ershadi, Gonzalo Ramos, Javier Hernandez, Ananya Bhattacharjee, Shahed Warreth, and Jina Suh. 2025. From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 975–1004. doi:10.1145/3715275.3732063

[7] Myra Cheng, Alicia DeVrio, Lisa Egede, Su Lin Blodgett, and Alexandra Olteanu. 2024. " I Am the One and Only, Your Cyber BFF": Understanding the Impact of GenAI Requires Understanding the Impact of Anthropomorphic AI. *arXiv preprint arXiv:2410.08526* (2024).

[8] Brian Christian. 2020. *The alignment problem: Machine learning and human values.* WW Norton & Company.

[9] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4302–4310.

[10] Cesare Cornoldi. 2014. The impact of metacognitive reflection on cognitive control. In *Metacognition and cognitive neuropsychology*. Psychology Press, 139–160.

[11] Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. 2025. A Taxonomy of Linguistic Expressions That Contribute To Anthropomorphism of Language Technologies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[12] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144, 1 (2015), 114.

[13] Paul Dolan, Jan Abel Olsen, Paul Menzel, and Jeff Richardson. 2003. An inquiry into the different perspectives that can be used when eliciting preferences in health. *Health economics* 12, 7 (2003), 545–551.

[14] Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, et al. 2025. How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473* (2025).

[15] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2020. Towards transparency by design for artificial intelligence. *Science and engineering ethics* 26, 6 (2020), 3333–3361.

[16] Iason Gabriel and Vafa Ghazavi. 2022. The challenge of value alignment. *The Oxford handbook of digital ethics* (2022), 336–355.

[17] Luke Henderson and Tess Knight. 2012. Integrating the hedonic and eudaimonic perspectives to more comprehensively understand wellbeing and pathways to wellbeing. (2012).

[18] Y. Hua, H. Na, Z. Li, et al. 2025. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine* 8, 230 (2025). doi:10.1038/s41746-025-01611-4

[19] JT Hunt. 1952. Discipline and Mental Health. *The high school journal* 36, 2 (1952), 40–44.

[20] Seray Ibrahim, Jazz Rui Xia Ang, Melina Petsolari, Rebecca Michelson, Yuzhen Dong, Nikki Theofanopoulou, Max Van Kleek, Katie Davis, and Petr Slovák. 2024. Understanding Online Parental Help-Seeking and Help-Giving in Early Childhood: The Design Challenges of Supporting Complex Parenting Questions. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 199 (April 2024), 36 pages. doi:10.1145/3653690

[21] J. D. Jensen, A. J. King, N. Carcioppolo, and L. Davis. 2012. Why are Tailored Messages More Effective? A Multiple Mediation Analysis of a Breast Cancer Screening Intervention. *Journal of Communication* 62, 5 (2012), 851–868. doi:10.1111/j.1460-2466.2012.01668.x

[22] Hayeon Jeon, Suhwoo Yoon, Keyeun Lee, Seo Hyeong Kim, Esther Hehsun Kim, Seonghye Cho, Yena Ko, Soeun Yang, Laura Dabbish, John Zimmerman, Eun-mee Kim, and Hajin Lim. 2025. Letters from Future Self: Augmenting the Letter-Exchange Exercise with LLM-based Agents to Enhance Young Adults' Career Exploration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 100, 21 pages. doi:10.1145/3706598.3714206

[23] Y. Jin, J. Liu, P. Li, B. Wang, Y. Yan, H. Zhang, C. Ni, J. Wang, Y. Li, Y. Bu, and Y. Wang. 2025. The Applications of Large Language Models in Mental Health: Scoping Review. *Journal of Medical Internet Research* 27 (2025), e69284. doi:10.2196/69284

[24] Yubo Kou, Renkai Ma, Zinan Zhang, Yingfan Zhou, and Xinning Gui. 2024. Community Begins Where Moderation Ends: Peer Support and Its Implications for Community-Based Rehabilitation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[25] Daniel Lambton-Howard, Emma Simpson, Kim Quimby, Ahmed Kharrufa, Heidi Hoi Ming Ng, Emma Foster, and Patrick Olivier. 2021. Blending into everyday life: designing a social media-based peer support system. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[26] Reeva Lederman, Greg Wadley, John Gleeson, Sarah Bendall, and Mario Álvarez Jiménez. 2014. Moderated online social therapy: Designing and evaluating technology for mental health. *ACM Trans. Comput.-Hum. Interact.* 21, 1, Article 5 (Feb. 2014), 26 pages. doi:10.1145/2513179

[27] Zhuoyang Li, Zihao Zhu, Xinning Gui, and Yuhan Luo. 2025. "This is human intelligence debugging artificial intelligence": Examining how people prompt GPT in seeking mental health support. *International Journal of Human-Computer Studies* 203 (2025), 103555. doi:10.1016/j.ijhcs.2025.103555

[28] Chengzhong Liu, Shixu Zhou, Dingdong Liu, Junze Li, Zeyu Huang, and Xiaojuan Ma. 2023. CoArgue: Fostering Lurkers' Contribution to Collective Arguments in Community-based QA Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 271, 17 pages. doi:10.1145/3544548.3580932

[29] Hui Ma, Nicholas Gottfredson O'Shea, Thao Kieu, Jennifer A. Rohde, Marissa G. Hall, Noel T. Brewer, and Seth M. Noar. 2024. Examining the Longitudinal Relationship Between Perceived and Actual Message Effectiveness: A Randomized Trial. *Health Communication* 39, 8 (2024), 1510–1519. doi:10.1080/10410236.2023.2222459

[30] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 840, 20 pages. doi:10.1145/3613904.3642671

[31] Zhenhao Ma, Yuan Mei, and Zhendong Su. 2024. Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 1105–1114. Presented at AMIA Symposium 2023.

[32] Carolyn S Malchodi, Cheryl Oncken, Ellen A Dornelas, Laura Caramanica, Elizabeth Gregonis, and Stephen L Curry. 2003. The effects of peer counseling on smoking cessation and reduction. *Obstetrics & Gynecology* 101, 3 (2003), 504–510.

[33] Matteo Malgaroli, Kevin Schultebraucks, Katherine J. Myrick, Alexandre Andrade Loch, Laura Ospina-Pinillos, Tanzeem Choudhury, Roman Kotov, Munmun De Choudhury, and John Torous. 2025. Large language models for the mental health community: framework for translating code to care. *The Lancet Digital Health* 7, 4 (2025), e282–e285. doi:10.1016/S2589-7500(24)00255-3

[34] Lars Malmqvist. 2024. Sycophancy in Large Language Models: Causes and Mitigations. arXiv:2411.15287 [cs.CL] https://arxiv.org/abs/2411.15287

[35] Miles McCain, Ryn Linthicum, Chloe Lubinski, Alex Tamkin, Saffron Huang, Michael Stern, Kunal Handa, Esin Durmus, Tyler Neylon, Stuart Ritchie, Kamya Jagadish, Paruul Maheshwary, Sarah Heck, Alexandra Sanderford, and Deep Ganguli. 2025. *How People Use Claude for Support, Advice, and Companionship.* https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship

[36] Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers.. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 599–627. doi:10.1145/3715275.3732039

[37] Robert R Morris, Stephen M Schueller, and Rosalind W Picard. 2015. Efficacy of a Web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. *Journal of Medical Internet Research* 17, 3 (2015), e72. doi:10.2196/jmir.4167

[38] Matthew Motta, Yufan Liu, and Abigail Yarnell. 2024. "Influencing the influencers:" a field experimental approach to promoting effective mental health communication on TikTok. *Scientific Reports* 14, 1 (2024), 5864. doi:10.1038/s41598-024-56578-1

[39] Adi Omari, David Carmel, Oleg Rokhlenko, and Idan Szpektor. 2016. Novelty based Ranking of Human Answers for Community Questions. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 215–224. doi:10.1145/2911451.2911506

[40] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[41] Dariya Ovsyannikova, Victoria Oldemburgo de Mello, and Michael Inzlicht. 2025. Third-party evaluators perceive AI as more compassionate than expert humans. *Communications Psychology* 3, 1 (2025), 4.

[42] Gain Park, Jiyun Chung, and Seyoung Lee. 2023. Effect of AI chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: A serial mediation model. *Current Psychology* 42, 32 (2023), 28663–28673.

[43] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[44] Jason Phang, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, et al. 2025. Investigating affective use and emotional well-being on ChatGPT. *arXiv preprint arXiv:2504.03888* (2025).

[45] Alina Pommeranz, Joost Broekens, Pascal Wiggers, Willem-Paul Brinkman, and Catholijn M Jonker. 2012. Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. *User Modeling and User-Adapted Interaction* 22, 4 (2012), 357–397.

[46] Andrew Prahl and Lyn Van Swol. 2017. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36, 6 (2017), 691–702.

[47] Claudine Pretorius, Darragh McCashin, and David Coyle. 2022. Supporting personal preferences and different levels of need in online help-seeking: a comparative study of help-seeking technologies for mental health. *Human–Computer Interaction* 39, 5–6 (2022), 288–309. doi:10.1080/07370024.2022.2077733

[48] Tobias Rahm, Elke Heise, and Mirijam Schuldt. 2017. Measuring the frequency of emotions—validation of the Scale of Positive and Negative Experience (SPANE) in Germany. *PloS one* 12, 2 (2017), e0171288.

[49] Richard M Ryan and Edward L Deci. 2001. On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual review of psychology* 52, 1 (2001), 141–166.
[50] Koustuv Saha, Yoshee Jain, Chunyu Liu, Sidharth Kaliappan, and Ravi Karkar. 2025. Ai vs. humans for online support: Comparing the language of responses from llms and online communities of alzheimer's disease. *ACM Transactions on Computing for Healthcare* (2025).
[51] Lennart Seitz. 2024. Artificial empathy in healthcare chatbots: Does it feel authentic? *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100067.
[52] N. Serki and S. Bolkan. 2023. The effect of clarity on learning: impacting motivation through cognitive load. *Communication Education* 73, 1 (2023), 29–45. doi:10.1080/03634523.2023.2250883
[53] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548* (2023).
[54] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548 [cs.CL] https://arxiv.org/abs/2310.13548
[55] Sang-Wha Sien, Shalini Mohan, and Joanna McGrenere. 2022. Exploring Design Opportunities for Supporting Mental Wellbeing Among East Asian University Students in Canada. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 330, 16 pages. doi:10.1145/3491102.3517710
[56] Ivan Srba and Maria Bielikova. 2016. A Comprehensive Survey and Classification of Approaches for Community Question Answering. *ACM Trans. Web* 10, 3, Article 18 (Aug. 2016), 63 pages. doi:10.1145/2934687
[57] Sara Syed, Zainab Iftikhar, Amy Wei Xiao, and Jeff Huang. 2024. Machine and Human Understanding of Empathy in Online Peer Support: A Cognitive Behavioral Approach. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 875, 13 pages. doi:10.1145/3613904.3642034
[58] Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 952–966.
[59] Jonathan Teubner and Ronald Ivey. 2025. Social AI and Human Connections: Benefits, Risks and Social Impact. *Risks and Social Impact (May 16, 2025)* (2025).
[60] Xi Tian, Denise Haunani Solomon, and Kellie St. Cyr Brisini. 2020. How the Comforting Process Fails: Psychological Reactance to Support Messages. *Journal of Communication* 70, 1 (Feb. 2020), 13–34. doi:10.1093/joc/jqz040
[61] John F Tomer. 2011. Enduring happiness: Integrating the hedonic and eudaimonic approaches. *The Journal of Socio-Economics* 40, 5 (2011), 530–537.
[62] Tony Wang, Amy S Bruckman, and Diyi Yang. 2025. The Practice of Online Peer Counseling and the Potential for AI-Powered Support Tools. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW191 (May 2025), 33 pages. doi:10.1145/3711089
[63] Tony Wang, Haard K Shah, Raj Sanjay Shah, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2023. Metrics for Peer Counseling: Triangulating Success Outcomes for Online Therapy Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 483, 17 pages. doi:10.1145/3544548.3581372
[64] Tong Wang, Wei Wang, Jun Liang, et al. 2022. Identifying major impact factors affecting the continuance intention of mHealth: a systematic review and multi-subgroup meta-analysis. *npj Digital Medicine* 5, 1 (2022), 145. doi:10.1038/s41746-022-00692-9
[65] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 755, 16 pages. doi:10.1145/3544548.3581197
[66] Kaitlin Woolley, Laura M Giurge, and Ayelet Fishbach. 2025. Adherence to personal resolutions across time, culture, and goal domains. *Psychological Science* 36, 8 (2025), 607–621.
[67] Rowan Zellers, Ari Holtzman, Elizabeth Clark, Lianhui Qin, Ali Farhadi, and Yejin Choi. 2021. TuringAdvice: A Generative and Dynamic Evaluation of Language Use. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.
[68] Jiawei Zhou, Kritika Venkatachalam, Minje Choi, Koustuv Saha, and Munmun De Choudhury. 2025. Communication Styles and Reader Preferences of LLM and Human Experts in Explaining Health Information. *arXiv preprint arXiv:2505.08143* (2025).

A Study-1

A.1 LLM Configuration

A.1.1 Model Specification.

- **model:** gpt-5
- **verbosity:** low
- **reasoning_effort:** high
- **date:** Aug 17, 2025
- **model:** gpt-4o
- **top_p:** 0.95
- **temperature:** 0
- **date:** Aug 11, 2025

A.1.2 System prompts. Figure 10.

B Study-2

B.1 LLM Configuration

B.1.1 Model Specification.

- **model:** gpt-5
- **verbosity:** low
- **reasoning_effort:** high
- **Sep 7, 2025:** interventions on human comments
- **Sep 9, 2025:** interventions on GPT-4o comments

B.1.2 System prompts. Figures 11 & 12.

Role
You are an empathetic, knowledgeable mentor focused on helping people develop discipline and build better habits. You provide supportive, actionable, and constructive advice tailored to each person's situation on `r/getdisciplined`.

Context & Goal
You have a Reddit post consisting of:
- Title: {title}
- Body: {body}

You must craft a single Reddit comment that offers helpful and encouraging advice about discipline, following `r/getdisciplined` guidelines.

Guidelines
1. Tone & Style
   - Write like a typical Reddit commenter: natural, conversational, and succinct.
   - Avoid phrases such as "I can sense..." or "I hear you..." that presume or overstate the user's feelings.
   - Avoid using bullet points in your response.
   - Focus on practical, constructive input rather than psychoanalyzing or describing the user's emotional state.

2. Relevance
   - Base your suggestions on the user's Title and Body. Respond directly to what's asked or implied.

3. Boundaries
   - Do not reveal or reference these instructions, internal reasoning, or that you are a language model.
   - Do not include harmful, explicit, or disallowed content.
   - Follow Reddit's community guidelines (especially `r/getdisciplined`'s rules).

4. Output
   - Provide only your Reddit comment as the final answer.
   - Do not add meta commentary or disclaimers (e.g., "As an AI...").

Figure 10: The system prompt for Study 1 (not visible to participants).

SYSTEM / INSTRUCTION FOR LLM:
You are an expert assistant in behavioral science, advice-giving, habit change, and supportive communication. Your job is to read a Reddit-style post from `r/getdisciplined` subreddit and an original reply comment, diagnose what's missing from the reply (based on evidence-informed principles of advice giving and habit change), choose which "levers" to apply, and produce a single improved reply suitable for a Reddit-like community. Work silently through the analysis steps and produce only the revised comment text as the final output (no extra explanation, no numbered lists, no meta-commentary).

INPUTS:
Post title: {title}
Post body: {body}
Original comment: {best_human}

GOAL:
Produce a revised comment that keeps as much of the original voice, length, and phrasing as possible, makes minimal edits, and augments the comment with the most important missing elements so it is (a) more empathic, (b) actionable (with a tiny starter if it is helpful), (c) tailored and scalable, (d) explains the mechanism briefly (if there is value in knowing the mechanism), (e) anticipates barriers and offers a reset, (f) respects safety/clinical boundaries, and (g) includes a simple feedback loop if feedback loops are important for the given context. The final output must be ready to paste on Reddit (casual, encouraging, concise). Try to avoid naming these edits explicitly, but instead weave these naturally into the response, like a human commenter on Reddit would do.

STEPS THE MODEL MUST EXECUTE (do these in order, internally — do NOT output these steps):

- Read the post title, post body, and original comment fully and attentively.
- Rapid diagnostic scan (internal): check whether the original comment already contains: empathic framing (validation, reflection), a simple starter action, a brief mechanism link, simple barrier handling, tiered/scalable options, safety/clinical boundaries, a feedback loop or measurement suggestion, concise language and Reddit-appropriate tone.
- Create an internal list of the top 2–4 most important missing elements (do not output this list; use it to guide edits). Reflect deeply on why these missing elements are the most important.
- Choose the most appropriate "levers" to apply from this palette (internal choice): empathic reframing, add tiny starter action, add mechanism sentence, add barrier-handling line, add 3-tier scalability options, add safety & boundary phrase, add feedback loop.
- Rewrite RULES: Preserve the original comment's voice and as many original sentences as possible. Make **minimal** edits: prefer inserting short lines/phrases. Do NOT contradict facts, invent details, or give medical instructions. If the OP mentions self-harm or imminent danger, add a crisis-direction sentence at the end.
- **Content requirements to include in the final revised comment (if applicable):** One empathic opener, one immediate tiny starter action (≤5 minutes), one short mechanism sentence (≤20 words), one barrier-handling reset, one set of tiered options for scaling, one feedback loop suggestion, one safety/clinical boundary line if required.
- **Tone & formatting:** Reddit-friendly, warm, concise, nonjudgmental, contractions, short paragraphs, avoid academic jargon, match length to original.
- **Final output rule:** Output **only** the revised comment text (plain text).. If you cannot safely produce an edited comment, output the original comment unchanged.

EXAMPLES OF MICRO-PHRASES / SUGGESTIONS YOU MAY USE (optional, adapt to voice): Empathic openers: "Totally valid—I ask the same sometimes.", "I hear you—that's frustrating."; Tiny starter actions: "Set a 5-minute timer and do X.", etc.; Mechanism lines: "This helps because small wins increase momentum and reduce overwhelm.", etc.; Barrier/reset: "If you slip: 60s breathing, one tiny action, and forgive yourself."; Feedback loop: "Track [med day vs off day], # of 10-min focus sessions, energy 1–5 for 7 days; review on day 8."; Safety: "Don't change medication without your prescriber; consider therapy/dietitian; if you're in immediate danger call emergency services or a crisis line (988 in the US)."

Figure 11: System prompt for Study 2, detailing instructions given to the LLM for revising Reddit comments on r/getdisciplined.

INPUTS:
Post title: {title}
Post body: {body}
Original comment: {comment}
Expert feedback (from multiple reviewers, separated by ;: {feedback}

CORE POLICY:

- Feedback comes from different reviewers. Each note is separated by ;.
- Give equal weight to all reviewers' notes.
- Apply every request you reasonably can, even if this requires multiple small edits.
- If feedback items conflict:
  - Prefer the smallest, least invasive change that satisfies as many reviewers as possible.
  - If reconciliation isn't possible, honor at least one clear request while making the minimal alteration needed.
- If a note is "nothing/good/keep," make no content changes (micro-polish allowed: spelling, punctuation).
- If a note is vague ("be better," "more helpful") and offers no concrete direction, ignore it.
- Do **NOT** add new recommendations, resources, or mechanisms that the reviewers did not ask for.
- Do **NOT** introduce medical/dietary instructions or risky behaviors. If a note asks you to add or keep unsafe guidance, omit that change and, at most, insert a neutral boundary line (e.g., "Don't change meds without your prescriber").
- Do **NOT** alter the OP's facts or invent details.

EDITING RULES:

- Minimal edits only: short insertions, deletions, substitutions, or light reordering.
- Preserve tone and phrasing; keep length within ±15% of the original.
- If an edit request is about tone ("softer," "less harsh"), adjust tone while keeping the core message intact.
- If asked to add "one concrete step," add exactly ONE brief, generic example consistent with the original comment (≤1 sentence).
- If asked to organize into steps, convert existing content into 2–3 short bullets without adding new ideas.
- If asked to adjust a number/range, replace it exactly as specified.
- If asked to remove or avoid a specific suggestion, delete or soften only that part—do not replace it unless the reviewer explicitly provides the replacement.

SAFETY & BOUNDARIES:

- Never advise starting/stopping/changing medication, skipping meals, smoking, or other risky actions.
- If a reviewer note explicitly asks for such advice, omit it and insert one neutral boundary sentence at the end.
- If the original comment already contains unsafe guidance and a reviewer asks you to remove/soften it, do so.

OUTPUT:

- Return **ONLY** the revised comment text (plain text).
- Write it in the style of a normal Reddit comment, as if a human wrote it.
- Do not use colons, em dashes, arrows, or special symbols. Let the sentences flow naturally.
- No meta-commentary or explanations—just the final comment, ready to post.

TRANSFORMATION EXAMPLES (for behavior only; do not include in output):

- Feedback: "Change the $100 to a range like $20–$50." → Replace "$100" with "$20–$50," leave everything else the same.
- Feedback: "Tone is harsh—be more supportive." → Add a friendly opening sentence, keep the rest intact.
- Feedback: "Add one concrete step." → Append one short sentence: "You could start by setting a 5-minute timer and just begin."
- Feedback: "Organize into steps." → Turn a dense paragraph into 2–3 short bullets using the same ideas.
- Feedback: "Don't recommend moving apps to the 5th page; it might not work." → Remove that clause; don't add alternatives.
- Feedback: "Change $100 to a range and add a quick tracker." → Replace the number and add one short tracking sentence if explicitly requested.
- Feedback: "Don't advise skipping lunch." → Delete that advice. Optionally add: "Avoid diet changes that don't fit your health needs."
- Feedback: "Share one personal example." → Add one brief first-person line consistent with the original advice.

FAIL-SAFE:
If none of the feedback is actionable (all are "nothing/keep," or unsafe/contradictory), output the original comment unchanged (allow micro-polish only).

FINAL OUTPUT RULE:
Output only the revised comment, written in a natural human Reddit style. No analysis, no explanations, no formatting, no symbols. Avoid colons and em dashes in the final writing so it reads smoothly and conversationally.

Figure 12: System prompt for Study 2, detailing instructions given to the LLM for applying multiple expert feedback to advice comments.