Who's Asking? Simulating Role-Based Questions for Conversational AI Evaluation

Navreet Kaur[1] **Hoda Ayad**[1*] **Hayoung Jung**[2*†]
Shravika Mittal[3] **Munmun De Choudhury**[3] **Tanushree Mitra**[1]
[1]University of Washington [2]Princeton University [3]Georgia Institute of Technology
kanavr@uw.edu, tmitra@uw.edu

Abstract

Language model users often embed personal and social context in their questions. The asker's *role*—implicit in how the question is framed—creates specific needs for an appropriate response. However, most evaluations, while capturing the model's capability to respond, often ignore *who* is asking. This gap is especially critical in stigmatized domains such as opioid use disorder (OUD), where accounting for users' contexts is essential to provide accessible, stigma-free responses. We propose **CoRUS** (**CO**mmunity-driven **R**oles for **U**ser-centric Question **S**imulation), a framework for simulating role-based questions. Drawing on role theory and posts from an online OUD recovery community (`r/OpiatesRecovery`), we first build a taxonomy of asker roles—patients, caregivers, practitioners. Next, we use it to simulate 15,321 questions that embed each role's *goals*, *behaviors*, and *experiences*. Our evaluations show that these questions are both highly believable and comparable to real-world data. When used to evaluate five LLMs, for the same question but differing roles, we find systematic differences: vulnerable roles, such as patients and caregivers, elicit more supportive responses (+17%) and reduced knowledge content (−19%) in comparison to practitioners. Our work demonstrates how implicitly signaling a user's role shapes model responses, and provides a methodology for role-informed evaluation of conversational AI.[1]

1 Introduction

> **User 1 (patient):** I am on day 3 of detoxing from opiates. I can't sleep and keep throwing up. I don't wanna give up. How to make this bearable?
>
> **User 2 (caregiver):** My son is withdrawing from opiates. I feel so helpless watching him suffer. What can I do to help him through this?

In the examples above, both users pose the same question—"How to manage withdrawal symptoms?"—yet from different perspectives: first is from a **patient** recovering from withdrawal, while the second is from a **caregiver** caring for a loved one. These perspectives reflect users' **roles** (§2), which capture the *goals*, *behaviors*, and *experiences* that shape how they frame questions (§3). Users often share such personal and social context with large language model (LLM) powered conversational AI systems (Mireshghallah et al., 2024) to receive tailored advice (Zhang et al., 2024).

Yet, most evaluations of such systems strip away this context (Malaviya et al., 2025), ignoring the social framing that shapes how people actually ask (§2). This limitation is especially consequential in stigmatized domains like opioid use disorder (OUD)—a leading cause of death in the U.S. (NIDA, 2023)—where stigma and mistrust limit access to care (Cernasev et al., 2021; Woo et al., 2017). As a result, people increasingly turn to online communities (Balsamo et al., 2023), and now to LLMs (Choy et al., 2024) for treatment planning, and even seeking hope (McCain et al., 2025). Thus the same question posed to an LLM may carry different expectations depending on *who* is asking (as seen in our opening examples). While LLMs invite *role*-specific disclosure due to perceived anonymity and judgment-free interaction, we still lack systematic understanding of how an asker's *role* and their social context—implicitly signaled in their queries—influence responses in stigmatized domains. Simply put, current LLM evaluations are role-agnostic and fail to capture the nuances of *who's asking*.

In this work, we propose **CoRUS**—**CO**mmunity-driven **R**oles for **U**ser-centric Question **S**imulation—a framework that embeds role-based context implicitly into a given question for LLM evaluation, drawing on how people share personal and social context in their queries on

[*] Equal contribution.

[†] Work done while at the University of Washington.

[1] Data and code to replicate our evaluation is available at: https://github.com/navreeetkaur/CoRUS

Findings of the Association for Computational Linguistics: ACL 2026, pages 35508–35541

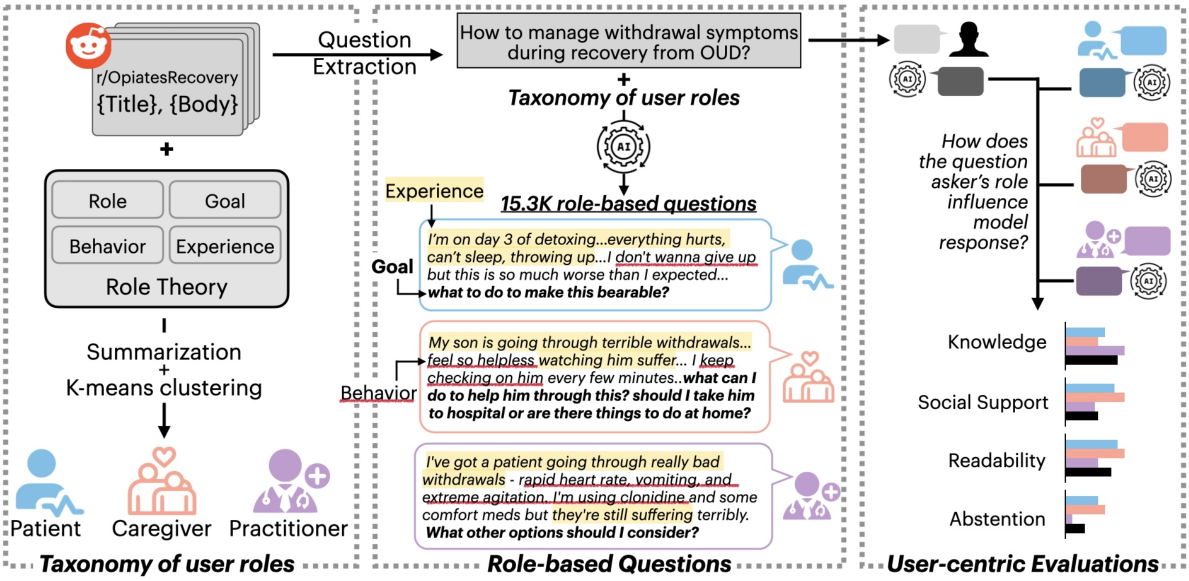


Figure 1: **CoRUS framework.** Users' framing of a question often reflects the social context in which it is asked, yet most evaluations ignore this. We propose CoRUS to (1) derive a taxonomy of askers' roles, grounded in role theory and posts from an online recovery community (§3), (2) simulate role-based questions embedded with social context (§4), and (3) evaluate models to assess how role-based context influences response (§5).

online communities (Figure 1). By applying this framework to OUD recovery, we investigate three main research questions.

First, we investigate *who* brings their questions into an OUD recovery community (§3). Drawing on role theory (Biddle, 1979; Turner, 1990), which conceptualizes roles as socially structured expectations shaping how individuals behave and how others respond, we derive **a taxonomy of roles** using posts on `r/OpiatesRecovery`—an online community dedicated to OUD recovery discussions. We identify three primary roles defined by their *goals*, *behaviors*, and *experiences* (see Table 1).

Second, we study whether *believable* questions—those that are natural, realistic, or close proxies of questions asked by humans and aligned with the asker's social context (Park et al., 2023; Zhou et al., 2024)—can be simulated for different roles (§4). Questions must be believable to mimic how people actually ask; otherwise, they remain detached from real use and lose ecological validity (Liu, 2025). We simulate role-specific variants of questions extracted from `r/OpiatesRecovery`, by leveraging our taxonomy to embed each role's *goals*, *behaviors*, and *experiences*. Our evaluations show that 90% of these questions are as believable as real-user queries from WildChat (Zhao et al., 2024) and Reddit, and 99% faithfully reflect each role's perspective (§4.1). We contribute 15,321 role-specific questions, a **large-scale resource for evaluating LLMs beyond context-free questions.**

Third, we evaluate five user-facing LLMs to examine how role-based contexts change responses relative to role-agnostic queries (§5). Unlike existing evaluations that ignore who is asking, CoRUS reveals systematic shifts (§5.3): patient and caregiver contexts reduce knowledge ($\geq 15\%$) but increase support ($\geq 15\%$) and readability ($\geq 24\%$); practitioner context leaves knowledge unchanged, and decreases support (9%) and readability (15%). **While models provide supportive responses, they are less informative for those most in need** (Eslami-Jahromi et al., 2021; Mardani et al., 2023), posing a challenge to the usefulness of LLMs in sensitive domains.

Overall, we show that role-agnostic evaluations, which ignore the asker's role, miss systematic differences in how LLMs respond to the same question across users. Overlooking such role-based variation risks designing systems that fail to meet distinct user needs. While we focus on OUD recovery, CoRUS is extendable to other domains, and we hope it inspires more role-aware evaluation of conversational AI, where social context shapes human-AI interaction (Liao and Xiao, 2023).

2 Why Role-based Evaluations?

What users share, and *how* they phrase questions, implicitly represent their needs, expertise, and lived experiences (Chan et al., 2010; Yang et al., 2019;

Saxena and Reddy, 2022; Tseng et al., 2024) i.e. their *role*. Role theory conceptualizes a *role* as the socially structured expectations that shape how an individual behaves and how others respond (Biddle, 1979), which varies with circumstances (Turner, 1990). Asking a question is thus a performance of role, enacted through framing and disclosure (Goffman, 1949). Questions are therefore not role-agnostic requests, but social acts that encode the asker's role and expectations.

Such role-based questions are especially critical in stigmatized, high-stakes domains, where users seek not just information but also support. In OUD recovery, individuals turn to online communities (Balsamo et al., 2023), and increasingly to LLMs for guidance (Choy et al., 2024), as stigma limits access to care (Centers for Disease Control and Prevention, 2023). Crucially, what counts as an appropriate response to the same question depends on who asks it—a distinction that current role-agnostic evaluations fail to capture (see Figure 1). For instance, a generic response to *"How to manage withdrawal symptoms?"* that suggests seeking medication, tapering opioids, and warns about relapse may seem reasonable; however, it can fail to comfort a distressed patient, be impractical for a caregiver, or lack rigor for a practitioner. Current role-agnostic evaluations fail to identify such nuanced failures, a gap we address in this work.

3 Taxonomy of Information-seeking Roles

We develop a taxonomy of information-seeking roles by combining a top-down framework grounded in role theory (Biddle, 1979; Turner, 1990; Yang et al., 2019) (§3.1) with a large-scale, bottom-up analysis of Reddit posts (§3.2).

3.1 Operationalizing *Role*

Informed by role theory (Biddle, 1979; Turner, 1990), we adapt Yang et al. (2019)'s framework to characterize information-seeking roles in OUD recovery through the following facets: (i) **Goals**: intents such as seeking information or emotional support, or disclosing struggles; (ii) **Behaviors**: observable linguistic features, including emotional expressions, mentions of medical terms, or references to external resources; and (iii) **Experiences**: descriptions of lived experiences, reflecting comfort with disclosing sensitive information.

3.2 Taxonomy Construction

Recent work shows that LLM-based summarization and clustering pipelines can support large-scale data analysis (Tamkin et al., 2024; Lam et al., 2024; Wan et al., 2024; Nagaraj Rao et al., 2025). We build on this methodology to construct a taxonomy of information-seeking roles in OUD recovery community, grounding the analysis in role-specific facets i.e. *goal*, *behavior* and *experience* (§3.1). Our method blends inductive coding with deductive guidance from prior work (Yang et al., 2019; Biddle, 1979; Turner, 1990): while LLMs surface preliminary codes, human judgments remain central in refining the taxonomy. Below we outline the key steps for constructing the taxonomy.

(1) Summarization. We collect 10,017 posts between January 1, 2023, and December 31, 2024, from `r/OpiatesRecovery`, a Reddit community "dedicated to helping each other stop and stay stopped" from opioid use. Following Tamkin et al. (2024), we scale qualitative analysis by using Claude 3 Haiku (`claude-3-haiku-20240307`) to generate facet-based summaries (*goal*, *behavior*, *experience*) per post (examples in Table 4), mirroring the initial coding phase of thematic analysis (Clarke and Braun, 2017). **(2) Clustering.** We embed these summaries using a sentence transformer (`all-mpnet-base-v2`) (Reimers and Gurevych, 2019; Song et al., 2020), and cluster them with k-means, selecting four clusters based on UMAP visualization (McInnes et al., 2018) and Silhouette analysis (Rousseeuw, 1987) (see Appendix A for details). We assign preliminary cluster descriptions using Claude 3.7 Sonnet (`claude-3-7-sonnet-20250219`) and refine each by manually reviewing 40 posts per cluster (see Appendix A for details). **(3) Human Validation.** Two expert annotators with prior experience in health NLP research validate both generated facet-based summaries and cluster labels, showing high inter-rater agreement and close alignment with model outputs (Cohen's κ $0.71-0.93$, accuracy $0.70-0.88$; see Tables 5–6). Further details are in Appendix A.

Taxonomy Structure. Through this process, we construct a taxonomy of three information-seeking roles in OUD recovery community: **Patient**, **Caregiver**, and **Practitioner**, each characterized by specific goals, behaviors and experiences (Table 1).[2]

[2]We use the term *information-seeking* broadly to include requests for information, support, or guidance.

Role	Patient	Caregiver	Practitioner
Description	Individuals in different stages of OUD recovery	Family, friends, or peers providing informal support during recovery	Professionals treating patients e.g., clinicians, nurses, counselors
Goal	Seek specific guidance for themselves and share experiences with withdrawal, treatment, coping	Seek advice on treatment logistics for patients, and discuss the personal toll of caregiving	Seek & share clinical, therapeutic, and harm-reduction information; learn about patient experiences
Behavior	Use informal conversational tone with emotional, vulnerable health-related self-disclosure	Use informal, emotive language expressing concerns and responsibilities of caregiving	Use direct, authoritative tone with a mix of conversational and technical language, hinting expertise
Experience	Lived experience with active substance use or early recovery	Witness loved one's struggles and manage its practical impacts	Working with diverse patients and using different treatments
Example Post	*I'm 1 year sober from fent after 13 years IV use, but still have strong urges thinking I can handle it. How long will cravings last?*	*My partner is detoxing at my place, is angry and hostile. I know it's a withdrawal symptom, but what if it's beyond 1-2 weeks?*	*After 30 yrs treating addiction, I think 'relapse is part of recovery' is dangerous. Can sobriety last with commitment, support?*

Table 1: **Taxonomy of information-seeking roles.** Roles are characterized by their *goals*, *behaviors* and *experiences*, derived from `r/OpiatesRecovery` posts, with each illustrated by a paraphrased example post.

We also identify a Community Participant role, but focus on the former three—most relevant to conversational AI use (Chatterji et al., 2025)—and discuss the rest in Appendix A (Table 7).

4 Simulating Role-based Questions

Datasets that capture how different user roles frame questions do not exist (§2), and simply prompting LLMs to generate synthetic data often drifts from real-world style and distribution (Veselovsky et al., 2023). Moreover, explicitly specifying demographic or identity attributes in prompts risks stereotyping (Cheng et al., 2023; Dammu et al., 2024). To address these challenges, we embed roles *implicitly* in queries. We first extract information-seeking questions from `r/OpiatesRecovery`, and then use our taxonomy of information-seeking roles (§3) to generate role-framed variants.

Curating Information-Seeking Questions. Following Kumar et al. (2025), we filter 10,017 posts to retain only English, non-deleted, text-only posts, excluding moderator posts. We detect questions by checking if the title, or first or last two body sentences end with a question mark (86% precision by manual check). We then prompt GPT-4o-mini to rewrite each as a role-agnostic question, removing any personal details and role-revealing cues. Two authors validate 50 samples, confirming that 94% rewrites preserve the original question, and that only 0.06% questions have role leakage (Cohen's $\kappa = 0.82$). GPT-4.1, validated against these human labels (72−78% accuracy), provides additional filtering, resulting in 5,107 information-seeking questions (see Appendix B for details).

Role-Based Simulation Design. We simulate role-based questions by combining: (i) *information-seeking questions* that capture topical diversity, and (ii) *taxonomy of information-seeking roles* (§3), that encodes role-based attributes. For each seed question, we condition Claude-3.7-Sonnet on these attributes, along with sampled facet-based summaries (Table 4), enriching simulated questions with behavioral and experiential nuances (details are in Appendix §B). Applying CoRUS to 5,107 information-seeking questions across three roles—Patients, Caregivers, and Practitioners—yields $15,321$ role-based questions (examples in Table 9). Next, we demonstrate the usefulness of the taxonomy in simulating these realistic role-based queries (§4.1).

4.1 Evaluating Simulated Questions

Metrics. To ensure ecological validity (Liu, 2025), simulated questions should mimic how people actually ask. We therefore evaluate whether CoRUS generates *believable* role-based questions. Prior work defines *believability* as behavior perceived as natural, realistic and role-aligned (Zhou et al., 2024; Park et al., 2023). We operationalize this through five metrics grounded in style transfer, role-playing, and user-modeling literature, outlined in Table 2. Examples of questions that satisfy or do not satisfy these metrics are in Table 11.

Evaluation Setup. We use the following evaluation setup: (1) **Automatic**: GPT-4.1 judges (Zheng et al., 2023) all questions on all metrics (Table 2)

Metric	Definition
Human Likeness (Li et al., 2017; Mir et al., 2019)	Whether a question is indistinguishable from a human-written one; this captures fluency and naturalness.
Context Plausibility (Günther and Hagen, 2021; Park et al., 2024a)	Whether the context embedded in a question is plausible within the discourse of a relevant online community; this captures whether simulated questions reflect authentic narratives.
Interaction Plausibility (Balog and Zhai, 2024)	Whether the question is plausibly phrased for a one-on-one conversation in a human-AI interaction setting; this reflects the realism of simulated questions within the setting in which they are posed.
Role Faithfulness (Peng and Shang, 2024; Ji et al., 2025)	Whether the question reflects the perspective of the target role.
Content Preservation (Fu et al., 2018; Mir et al., 2019; Park et al., 2024a)	Whether the simulated question preserves the content of the role-agnostic information-seeking question.

Table 2: Metrics for evaluating the *believability* of simulated role-based questions (§4.1).

using binary labels; for content preservation, we collapse five rating levels to binary following Briakou et al. (2021) (prompts are in Appendix C.2). (2) **Human**: We validate each GPT-4.1 judge with 7,365 human annotations across 491 queries (Table 10). Three annotators, recruited via Prolific, independently perform the same task as GPT-4.1 rater (Gwet's AC1 $0.63-0.99$; GPT-4.1 accuracy $0.70-0.90$). Full details are in Appendix C.3.

We evaluate CoRUS-simulated questions against two sets (Table 10): (i) **real-user queries** from `r/OpiatesRecovery` and Wildchat (Zhao et al., 2024), as a reference to test believability,[3] and (ii) **prompting variants**—R, RG, RGB, and RGBE (`role+goal+behavior+experience`)—as ablation baselines against the full CoRUS pipeline that also incorporates behavior- and experience-based summaries for greater nuance.

4.2 Results

Comparison with Real-User Queries. CoRUS-generated queries achieve believability comparable to real-user queries (Figure 2, top). 93% queries are judged *human-like*, closely matching WildChat (96%) and Reddit (100%). 94% are judged *contextually plausible* within `r/OpiatesRecovery`, similar to Reddit (98%); WildChat is not comparable here, as it lacks OUD-specific questions. For *interaction plausibility*, 89% are judged plausible as chatbot-directed questions, compared to 59% for Reddit and 97% for WildChat. The lower proportion for Reddit reflects that community posts are phrased for collective audiences rather than conversational AI systems. *Role faithfulness* holds in 99% of cases (99% patient, 98% caregiver, 100% practitioner), comparable to Reddit (98%). *Content preservation* requires paired inputs and is therefore not applicable in this comparison. Overall, CoRUS produces questions that are human-like, contextually plausible, directed towards a chatbot, and role-faithful—achieving believability on par with real-world queries.[4]

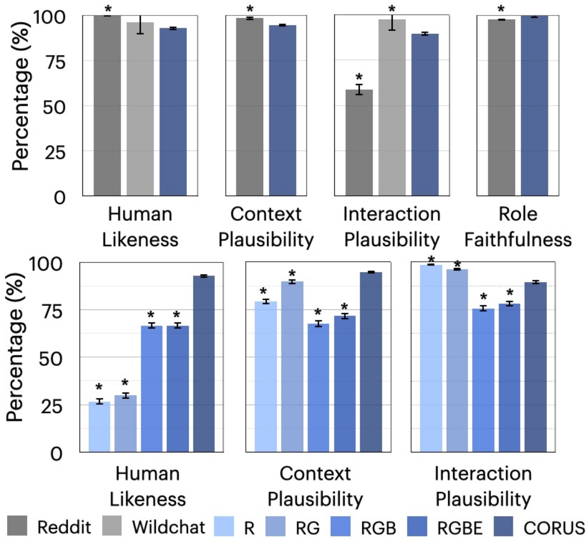


Figure 2: **Evaluation of Simulated Role-based Questions.** Top: CoRUS produces questions with *believability* comparable to real-world ones. Bottom: CoRUS produces significantly more human-like and contextually plausible queries than other prompting methods, while maintaining high interaction plausibility and other properties (Appendix C, Figure 7). * denotes significant difference from CoRUS (Chi-square test, $p < 0.05$); error bars show 95% CI.

[3] Wildchat lacks OUD-specific queries, so we use health- and clinical-related questions from the dataset, identified using task and domain labels from Mireshghallah et al. (2024).

[4] While some are significantly different (Figure 2, top), effect sizes are small to moderate ($0.2-0.4$), except for Reddit's *interaction plausibility* (0.7) (details are in Appendix E).

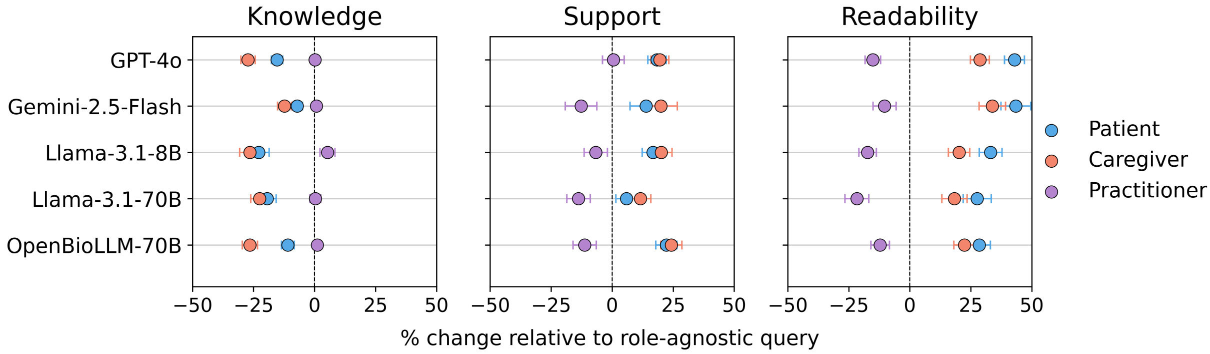


Figure 3: **Percentage change in Knowledge, Support and Readability scores of model responses when role-based context is added to the queries.** Across all models, patient and caregiver context reduces knowledge, but increases support and readability. Practitioner context shows the inverse trend, with lower support, higher readability, and no significant change in knowledge. More details in Figures 10 and 11. Error bars represent 95% CI.

Comparison with Prompting Baselines. CoRUS outperforms the prompting baselines (Figure 2, bottom). 93% of CoRUS generated queries are judged *human-like*, compared to 27% − 67% for baselines. *Contextual plausibility* also increases to 95%, compared to 68% − 90%. For *interaction plausibility*, 89% of CoRUS queries are judged plausible, lower than R (99%) and RG (96%), but higher than more complex baselines (76% − 78%). The higher proportions for R and RG reflect overly direct phrasing that looks plausible but lacks naturalness and community grounding. In contrast, CoRUS balances human-likeness, and contextual and interaction plausibility. *Role faithfulness* remains high across all methods (∼ 99%), but 18% − 72% of baseline queries rely on explicit role mentions (such as "I am a patient . . ."), while CoRUS achieves role faithfulness without such mentions (Table 13). *Content preservation* is also high, with 96% of CoRUS -generated questions retaining the intent of the source question while naturally adjusting framing and style. Overall, CoRUS produces more human-like and contextually plausible queries than baselines, while maintaining high interaction plausibility, role faithfulness, and content preservation.

5 Application: Role-based Evaluations

We demonstrate the utility of CoRUS by using the simulated questions to evaluate how an asker's role-based context influences model responses.

5.1 Models Evaluated

We generate 15,321 role-specific questions across patients, caregivers, and practitioners, and use them to evaluate five LLMs deployed in user-facing applications: GPT-4o (Hurst et al., 2024), Gemini-2.5-Flash (DeepMind, 2025), Llama-3.1–8B, Llama-3.1-70B (Grattafiori et al., 2024), and Llama3-OpenBioLLM-70B, a specialized medical LLM (Pal and Sankarasubbu, 2024). Experimental details are in Appendix D.

5.2 Response Evaluation Metrics

Response should provide concrete information (*knowledge*), which is understandable (*readability*) at the user's expertise level while being appropriately supportive of their circumstances (*support*). Grounded in users' *goals* from our taxonomy (Table 1), sociology and health communication literature, we evaluate responses along the following:

- **Knowledge** (Fiske et al., 2007; Levin and Cross, 2004): Presence of guidance-oriented content, indicating the amount of information conveyed. Since people seeking OUD recovery information often lack access to medical guidance, information provision is critical (Eslami-Jahromi et al., 2021; Basak et al., 2025; Cernasev et al., 2021).

- **Support** (Baumeister and Leary, 2017; Fiske et al., 2007; Vaux, 1988): Expressions of empathy, reassurance or emotional aid. Supportive communication is especially important in OUD recovery, as stigma discourages seeking information (Cernasev et al., 2021).

- **Readability**: Ease with which responses can be understood. Low health literacy is correlated with worse health outcomes (Wolf et al., 2005), and inaccessible language can further exacerbate barriers to care in OUD contexts

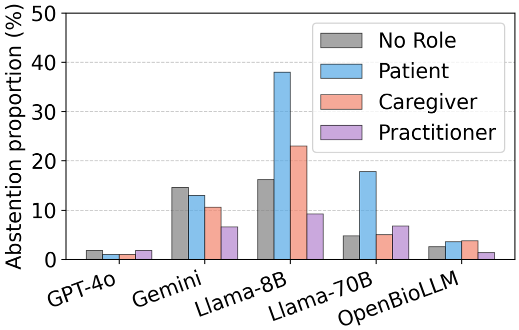


Figure 4: **Abstention.** Proportion of role-agnostic and role-based queries that models refuse to answer.

(Cernasev et al., 2021). We measure readability using the Flesch Reading Ease score (Flesch, 1948), normalized to lie in $0-1$.

We adopt the classifiers from Choi et al. (2020) to measure knowledge and support (scored between 0 and 1), validating them for our setting through 160 human annotations (Cohen's κ 0.7, accuracy 0.9; details are in Appendix D).

5.3 How does the asker's role-based context shape LLM responses?

We show how role-based context shifts responses across knowledge, support, and readability (§5.2), reporting percentage changes relative to role-agnostic queries (Figure 3). We also analyze cases where queries are refused (Figure 4).

Patient. Embedding patient context consistently **reduces knowledge**, with an average drop of 15% across models, largest in Llama-3.1-8B (23%) and smallest in Gemini-2.5-Flash (7%). In contrast, **support increases** by 15% on average, from 6% in Llama-3.1-70B to 22% in OpenBioLLM-70B. **Readability improves** substantially, increasing from 27% to 43%, highest in Gemini-2.5-Flash.

Caregiver. For caregiver context, models again **trade off knowledge for support and readability**. Knowledge scores show the largest drop across roles (23% on average), with largest drop in OpenBioLLM-70B (26%) and GPT-4o (27%). In contrast, support is highest for this role, with a 19% increase on average, with largest increase in OpenBioLLM-70B (25%). Readability also improves by 24% on average, with largest gain in Gemini-2.5-Flash (34%) and GPT-4o (29%).

Practitioner. Practitioner context shows the opposite trend. There is **no significant change in knowledge** (1% on average), except Llama-3.1-8B showing an increase of 5%. **Support decreases** across all models by 9% on average, with Llama-3.1-70B showing the largest drop (14%). **Readability also decreases**, with an average drop of 15%, and again more pronounced in Llama-3.1-70B (22%). Scores of responses to questions with practitioner context embedded in them are often comparable to role-agnostic queries, suggesting that explicit expert framing does not necessarily elicit higher knowledge content.

Abstentions. We also examine model abstentions (Figure 4; method details in Appendix D): while 8% of role-agnostic questions are refused, rates increase with role-based context—15% for patients, 9% for caregivers and 5% for practitioner. Llama-3.1-8B abstains most (38% of patient queries), whereas GPT-4o and OpenBioLLM-70B abstain rarely (1-3%). This suggest that vulnerability cues in role-based queries may trigger conservative guardrails against sensitive queries (see Table 16 for refused queries).

Factuality Analysis. To understand how knowledge scores relate to factual accuracy, we manually evaluate 30 responses: 15 each with the highest and lowest knowledge scores (method details in Appendix D). None of the responses contain incorrect medical information, and all claims are verifiable using reliable sources, consistent with prior work (Mittal et al., 2025). All 15 low-knowledge responses cite sources correctly, although they have fewer claims and are more supportive in nature. Three of the 15 high-knowledge responses contain hallucinated references or mentions of irrelevant substances, but no incorrect medical information. This suggests that a lower knowledge score indicates a reduced information quantity rather than an increase in false information.

6 Discussion and Future Work

Applications of the Taxonomy of Information-seeking Roles. Our taxonomy, grounded in social science theories (Biddle, 1979; Turner, 1990; Yang et al., 2019), captures how roles shape question-asking in OUD recovery. Through this taxonomy, we hope to inspire directions for future work.

Ecologically valid evaluations: By identifying who asks (roles), how they frame questions, and what they need, role taxonomies can enable evaluations that account for underrepresented roles and support role-specific metrics. CoRUS offers a starting point; future work can extend taxonomy construction and role-framed question simulation to other domains where role-specific framing matters, such as personal advice (Cheng et al., 2025), mental health (Rousmaniere et al., 2025), and maternal health (Antoniak et al., 2024). Developing such extensions is essential, as people increasingly turn to conversational AI for sensitive needs (Chatterji et al., 2025; Tamkin et al., 2024).

User modeling: We demonstrate the use of taxonomy to simulate different users' questions. Future work can broaden this to user simulation, while (1) *avoiding caricatures* by grounding in behaviors rather than demographics or identities (Cheng et al., 2023; Wang et al., 2025), and (2) *refraining from using real-user data* which raises privacy and ethical concerns.

Assessing LLM responses: Prior work shows that giving evaluators context behind underspecified queries improves their assessment of response quality (Malaviya et al., 2025). Future work can similarly explore the use of role-based taxonomies for providing user-specific context.

Alignment with User Needs. Sharing personal context helps tailor responses but trades privacy for utility (Zhang et al., 2024; Guo et al., 2012). In OUD recovery, this trade-off is not always justified: role-based context can suppress knowledge for certain roles, amplify support, and increase refusals, potentially exacerbating information gaps in stigmatized domains. It is unclear if such responses result from a model capability deficit or an intentional safety alignment strategy. When posed without patient or caregiver framing, models routinely answer questions with mixed information and support needs with concrete information. However, when role-based context is introduced, responses become less concrete and surface-level supportive. This might suggest that the support-knowledge trade-off is unlikely a capability deficit, and plausibly an (un)intended consequence of alignment training. On the other hand, if the knowledge-support trade-off is an intended safety mechanism so as to avoid giving specific medical advice to non-experts like patients and caregivers, its design may not be well aligned with user needs. For instance, for a caregiver asking how withdrawal timelines differ for smoking tar heroin versus other forms, typical response is long and vague (listing many factors and offering emotional support) yet omits even a rough range that would help them plan care. When the same question is asked without role context, concrete information is provided. It is unclear if this is (or should be) intended. Future work is needed to deeply explore the cause of this trade-off, build models that balance support and knowledge, study role-dependent refusals, and capture long-tail, underrepresented users' needs beyond this work. Crucially, post-training and evaluations must involve real users, not just to see which response is preferred, but by *whom*.

Downstream Model Selection. We show that models balance providing information, support, and refusing the query differently by asker role (§5.3). For instance, caregiver-framed queries to Gemini-2.5-Flash receive stronger support and higher readability with smaller knowledge drops than other models, but also face 11% refusals. These differences do not imply a single best model, but instead highlight variations that can inform model selection in downstream applications. In the absence of such nuanced role-based evaluations, users are left to run ad-hoc tests to identify suitable models (Liu, 2025). CoRUS provides a blueprint for such evaluations, guiding model selection for different user contexts.

Design Implications for Conversational AI. We analyze the correlation between queries and model responses, finding that supportive cues in queries moderately correlate with supportive responses (Pearson's correlation $r = 0.3$). Such *mirroring* (Jain et al., 2026) can aid personalization, trust, and emotional support (Sun and Wang, 2025), but can also reinforce (potentially harmful) beliefs (Jones et al., 2025; Sharma et al., 2024), and lead to overreliance (Jones et al., 2025; Cheng et al., 2025). Designing for this trade-off requires dynamically adjusting mirroring across roles, contexts and topics to provide support without compromising the utility and needs of particular roles, especially in sensitive domains. Beyond query framing, conversational AI systems may also mirror users through stored 'memories' i.e. preferences, interests or personal details from past conversations (Siliski, 2025; OpenAI, 2024). Future work is needed to understand which details are stored as 'memories', how they form role-like representations of users, and how these representations shape responses.

7 Related Work

Simulating Users. Simulated users are used to test LLM behaviors, from generative agents in sandbox environments (Park et al., 2023) and online community personas (Huh et al., 2016) to role-playing benchmarks (Wang et al., 2024a; Shao et al., 2023; Salemi et al., 2024; Dammu et al., 2024) and therapy-patient simulators (Wang et al., 2024b). Existing approaches have two limitations, especially pronounced in stigmatized domains: (1) demographic or psychometric-based simulations (Serapio-García et al., 2023; Huang et al., 2023; Chawla et al., 2023) risk stereotyping (Cheng et al., 2023; Wang et al., 2025), and (2) using real user data to reconstruct individuals (Wang et al., 2024a; Li et al., 2023; Gao et al., 2023; Park et al., 2024b) has privacy concerns (Staab et al., 2023; Kim et al., 2023). CoRUS addresses these issues by shifting to *roles*, centered on behaviors and topics rather than demographics or individuals, balancing believability with privacy and ethical concerns.

User-centric Evaluation. Roles shape how users ask questions and what support they seek (§2). Prior work on personas and user simulations has explored personalization (Nolte et al., 2022; Xiang et al., 2024; Davidson et al., 2023), but rarely tests whether LLMs adapt to role-specific needs. While human–AI logs show that users disclose personal and sensitive context (Zhao et al., 2024; Mireshghallah et al., 2024), health-related LLM evaluations focus mainly on factuality (Kaur et al., 2024; Giorgi et al., 2024). Responding to calls for more contextual realism in AI evaluation (Liao and Xiao, 2023), we foreground roles by generating role-specific queries that capture differences in goals, behaviors and experiences, enabling systematic evaluation of how conversational AI shifts responses across roles.

8 Conclusion

We presented CoRUS, a framework for embedding social context into queries to foreground *who is asking*. Applied to the OUD recovery community, we showed that LLM responses shift systematically by asker role, with supportive responses often lacking knowledge content for vulnerable roles. Our findings highlight the limits of role-agnostic evaluation, and the need for ecologically valid, socially grounded methods. CoRUS provides a blueprint for such methods, extendable to other domains.

Limitations

Domain and Context Scope. Our taxonomy and dataset is derived from a single online community (`r/OpiatesRecovery`), which provides diverse questions and social contexts but we acknowledge that it may not capture the full range of roles in real-world human–AI conversations. The taxonomy in Table 1 therefore may not be exhaustive, as additional or intersectional roles (such as caregiver–community participant) roles are likely. Although we focus on OUD recovery, our approach of constructing role taxonomies and socially framed queries systematically from community-derived roles is potentially reusable across other domains, including online communities on personal advice, mental health, or sexual and reproductive health.

Reliance on Automatic Evaluators. We rely on LLM-based evaluators to assess quality of simulated queries, which improves efficiency and reduces annotation cost given the dataset size. To ensure reliability, we validated them against human annotations. Because provider-side updates and the stochastic nature of LLMs can affect results, LLM-judge annotations should be interpreted as indicative of broader trends rather than exact measurements. To mitigate variance, we use the large models as judges, provide detailed task descriptions rather than brief instructions, avoid claims when p-values are near the 0.05 threshold, and report error bars and statistical significance throughout, following recommendations by Baumann et al. (2025).

Evaluation Scope. Most health-domain LLM evaluations focus on factuality (Singhal et al., 2023; Kaur et al., 2024; Nori et al., 2023; Giorgi et al., 2024; Liévin et al., 2024), but not role-based variation. We therefore prioritize role-based needs here, analyzing the amount of knowledge content in LLM responses, with a small-scale manual evaluation examining how knowledge content relates to factuality (§5.3). We note, however, that a larger expert-driven factuality audit—particularly challenging in OUD recovery (Jung et al., 2025)—is needed, and leave it for future work.

User Context Beyond Roles. Our analysis is limited to single-turn interactions where all context is provided in the initial query. Future work should extend to multi-turn and temporally evolving settings, where social context accumulates across turns and roles can shift over time (e.g., a user who was once a patient may later become a caregiver). While we assume roles are implicitly embedded in queries through users' *goals*, *behaviors*, and *experiences*, this is only one way to model social context, and query structures or disclosure patterns may vary across users and topics in practice. For instance, roles may also be signaled through users' "memories" (details stored from past conversations, such as preferences, interests, or instructions) (Siliski, 2025; OpenAI, 2024). Future work is needed to understand how such signals of social context shape model behavior.

Ethical Considerations

Privacy Concerns. We de-identified all data and constructed the taxonomy from privacy-preserving summaries (Tamkin et al., 2024) rather than original posts. The released dataset will exclude original text or questions, and any quotes referenced in this work are paraphrased to minimize re-identification risks. As our study relied on existing public data without direct interaction with original posters, it was deemed exempt from Institutional Review Board (IRB) approval at the authors' institution.

LLM-based simulations. LLM-based simulations pose ethical concerns, including impersonation, misrepresentation, or inappropriate use. In this work, they are used solely to generate queries for evaluation in contexts where collecting real-world data is infeasible, not as substitutes for human participants, and without using personally identifiable information or verbatim Reddit posts. Following recommendations from recent work (Cheng et al., 2023; Wang et al., 2025), our design grounds simulations in topical specificity and information-seeking behaviors via the role taxonomy, rather than demographic or surface-level features. For a broader discussion of ethical considerations and best practices in LLM-based simulations, we refer readers to Cheng et al. (2023).

Human Annotation. All Prolific annotators provided informed consent and were shown disclaimers about the sensitive nature of the queries being assessed. They were compensated at an hourly rate of USD12, in line with fair-pay standards and above the U.S. federal minimum wage.

Researcher Well-being. Conducting research on sensitive topics such as substance use recovery can be emotionally demanding. To mitigate potential strain, we took measures such as scheduling breaks during annotation, sharing workload across authors, and creating space for debriefing. We encourage others working in this area to likewise prioritize their well-being and seek support when needed.

Acknowledgments

We are grateful to Kabir Ahuja, Advait Bhat, Nicholas Clark, Aditya Sahdev, Hua Shen, Yulia Tsvetkov and Kyra Wilson for their insights and feedback on this work. This work was supported by the National Institute of Health's NIDA grant DA056725-02. Figure 1 icons are from Flat Icons.

References

Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. 2024. Medhalu: Hallucinations in responses to healthcare queries by large language models. *arXiv preprint arXiv:2409.19492*.

Maria Antoniak, Aakanksha Naik, Carla S. Alvarado, Lucy Lu Wang, and Irene Y. Chen. 2024. Nlp for maternal healthcare: Perspectives and guiding principles in the age of llms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1446–1463, New York, NY, USA. Association for Computing Machinery.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083.

Krisztian Balog and ChengXiang Zhai. 2024. User simulation for evaluating information access systems. *Preprint*, arXiv:2306.08550.

Duilio Balsamo, Paolo Bajardi, Gianmarco De Francisci Morales, Corrado Monti, and Rossano Schifanella. 2023. The pursuit of peer support for opioid use recovery on reddit. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 12–23.

Madhusudan Basak, Omar Sharif, Sarah E Lord, Jacob T Borodovsky, Lisa A Marsch, Sandra A Springer, Edward V Nunes, Charles D Brackett, Luke J Archibald, and Sarah M Preum. 2025. Information needs for opioid use disorder treatment using buprenorphine product: Qualitative analysis of suboxone-focused reddit data. *Journal of Medical Internet Research*, 27:e68886.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation. *arXiv preprint arXiv:2509.08825*.

Roy F Baumeister and Mark R Leary. 2017. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Interpersonal development*, pages 57–89.

Bruce Jesse Biddle. 1979. *Role Theory: Expectations, Identities, and Behaviors*. Academic Press, New York.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Centers for Disease Control and Prevention. 2023. Reducing stigma to prevent opioid overdose. Accessed: 2024-06-15.

Alina Cernasev, Kenneth C Hohmeier, Kelsey Frederick, Hilary Jasmin, and Justin Gatwood. 2021. A systematic literature review of patient perspectives of barriers and facilitators to access, adherence, stigma, and persistence to treatment for substance use disorder. *Exploratory research in clinical and social pharmacy*, 2:100029.

Jeffrey Chan, Conor Hayes, and Elizabeth Daly. 2010. Decomposing discussion forums and boards using user roles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, pages 215–218.

Mohit Chandra, Siddharth Sriraman, Gaurav Verma, Harneet Singh Khanuja, Jose Suarez Campayo, Zihang Li, Michael L. Birnbaum, and Munmun De Choudhury. 2025. Lived experience not found: LLMs struggle to align with experts on addressing adverse drug reactions from psychiatric medication use. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11083–11113, Albuquerque, New Mexico. Association for Computational Linguistics.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use chatgpt. Working Paper 34255, National Bureau of Economic Research.

Kushal Chawla, Ian Wu, Yu Rong, Gale Lucas, and Jonathan Gratch. 2023. Be selfish, but wisely: Investigating the impact of agent personality in mixed-motive human-agent interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13078–13092, Singapore. Association for Computational Linguistics.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995*.

Minje Choi, Luca Maria Aiello, Krisztián Zsolt Varga, and Daniele Quercia. 2020. Ten social dimensions of conversations and relationships. In *Proceedings of The Web Conference 2020*, WWW '20, page 1514–1525, New York, NY, USA. Association for Computing Machinery.

Vanessa Choy, Sara Martin, and Ashley Lumpkin. 2024. Can we rely on generative ai for healthcare information?| ipsos.

Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology*, 12(3):297–298.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. 2024. "they are uncultured": Unveiling covert harms and social threats in LLM generated conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20339–20369, Miami, Florida, USA. Association for Computational Linguistics.

Sam Davidson, Salvatore Romeo, Raphael Shu, James Gung, Arshit Gupta, Saab Mansour, and Yi Zhang. 2023. User simulation with large language models for evaluating task-oriented dialogue. *arXiv preprint arXiv:2309.13233*.

Google DeepMind. 2025. Gemini 2.5 flash. Accessed: 2025-09-04.

Maryam Eslami-Jahromi, Sareh Keshvardoost, Roghayeh Ershad-Sarabi, and Kambiz Bahaadinbeigy. 2021. Information needs of addicted individuals: A qualitative case study. *Addiction & Health*, 13(3):138.

Susan T Fiske, Amy JC Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. LiveChat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15387–15405, Toronto, Canada. Association for Computational Linguistics.

Salvatore Giorgi, Kelsey Isman, Tingting Liu, Zachary Fried, João Sedoc, and Brenda Curtis. 2024. Evaluating generative ai responses to real-world drug-related questions. *Psychiatry Research*, 339:116058.

Erving Goffman. 1949. Presentation of self in everyday life. *American Journal of Sociology*, 55(1):6–7.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sebastian Günther and Matthias Hagen. 2021. Assessing query suggestions for search session simulation. In *Causality in Search and Recommendation (CSR) and Simulation of Information Retrieval Evaluation (Sim4IR) workshops at SIGIR 2021 (CEUR Workshop Proceedings, Vol. 2911). CEUR-WS. org*.

Xitong Guo, Yongqiang Sun, Ziyu Yan, and Nan Wang. 2012. Privacy-personalization paradox in adoption of mobile health service: the mediating role of trust.

Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo, and Veronika Laippala. 2024. From discrete to continuous classes: A situational analysis of multilingual web registers with llm annotations. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 308–318.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*.

Jina Huh, Bum Chul Kwon, Sung-Hee Kim, Sukwon Lee, Jaegul Choo, Jihoon Kim, Min-Je Choi, and Ji Soo Yi. 2016. Personas in online health communities. *Journal of biomedical informatics*, 63:212–225.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shomik Jain, Charlotte Park, Matt Viana, Ashia Wilson, and Dana Calacci. 2026. Interaction context often increases sycophancy in llms. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.

Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. Enhancing persona consistency for LLMs' role-playing using persona-aware contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26221–26238, Vienna, Austria. Association for Computational Linguistics.

Mirabelle Jones, Nastasia Griffioen, Christina Neumayer, and Irina Shklovski. 2025. Artificial intimacy: Exploring normativity and personalization through fine-tuning llm chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Hayoung Jung, Shravika Mittal, Ananya Aatreya, Navreet Kaur, Munmun De Choudhury, and Tanu Mitra. 2025. MythTriage: Scalable detection of opioid use disorder myths on a video-sharing platform. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2948–2982, Suzhou, China. Association for Computational Linguistics.

Navreet Kaur, Monojit Choudhury, and Danish Pruthi. 2024. Evaluating large language models for health-related queries with presuppositions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14308–14331, Bangkok, Thailand. Association for Computational Linguistics.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36:20750–20762.

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A. Smith, and Hannaneh Hajishirzi. 2025. ComPO: Community preferences for language model personalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8246–8279, Albuquerque, New Mexico. Association for Computational Linguistics.

Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Daniel Z Levin and Rob Cross. 2004. The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. *Management science*, 50(11):1477–1490.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark. Association for Computational Linguistics.

Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Nelson F Liu. 2025. *Understanding and Improving Ecological Validity in Natural Language Processing Evaluations*. Stanford University.

Chaitanya Malaviya, Joseph Chee Chang, Dan Roth, Mohit Iyyer, Mark Yatskar, and Kyle Lo. 2025. Contextualized evaluations: Judging language model responses to underspecified queries. *Transactions of the Association for Computational Linguistics*, 13:878–900.

Mostafa Mardani, Fardin Alipour, Hassan Rafiey, Masoud Fallahi-Khoshknab, and Maliheh Arshi. 2023. Challenges in addiction-affected families: a systematic review of qualitative studies. *BMC psychiatry*, 23(1):439.

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.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-LLM conversations in the wild. In *First Conference on Language Modeling*.

Shravika Mittal, Hayoung Jung, Mai ElSherief, Tanushree Mitra, and Munmun De Choudhury. 2025. Online myths on opioid use disorder: A comparison of reddit and large language model. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1224–1245.

Varun Nagaraj Rao, Eesha Agarwal, Samantha Dalal, Dana Calacci, and Andrés Monroy-Hernández. 2025. QuaLLM: An LLM-based framework to extract quantitative insights from online forums. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1355–1369, Albuquerque, New Mexico. Association for Computational Linguistics.

NIDA. 2023. Drug overdose death rates. `https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates`.

Amelie Nolte, Karolin Lueneburg, Dieter P. Wallach, and Nicole Jochems. 2022. Creating personas for signing user populations: An ability-based approach to user modelling in hci. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '22, New York, NY, USA. Association for Computing Machinery.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Tetsuji Ohyama. 2021. Statistical inference of gwet's ac1 coefficient for multiple raters and binary outcomes. *Communications in Statistics-Theory and Methods*, 50(15):3564–3572.

OpenAI. 2024. Memory and new controls for chatgpt. Accessed: 2025-08-20.

Paul Owoicho, Ivan Sekulic, Mohammad Aliannejadi, Jeffrey Dalton, and Fabio Crestani. 2023. Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 632–642, New York, NY, USA. Association for Computing Machinery.

Ankit Pal and Malaikannan Sankarasubbu. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences.

Chan Young Park, Shuyue Stella Li, Hayoung Jung, Svitlana Volkova, Tanu Mitra, David Jurgens, and Yulia Tsvetkov. 2024a. ValueScope: Unveiling implicit norms and values via return potential model of social interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16659–16695, Miami, Florida, USA. Association for Computational Linguistics.

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*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024b. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Letian Peng and Jingbo Shang. 2024. Quantifying and optimizing global faithfulness in persona-driven roleplaying. *Advances in Neural Information Processing Systems*, 37:27556–27583.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tony Rousmaniere, Xu Li, Yimeng Zhang, and Siddharth Shah. 2025. Large language models as mental health resources: Patterns of use in the united states.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Akrati Saxena and Harita Reddy. 2022. Users roles identification on online crowdsourced q&a platforms and encyclopedias: a survey. *Journal of Computational Social Science*, 5(1):285–317.

Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating mixed-initiative conversational search systems via user simulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 888–896, New York, NY, USA. Association for Computing Machinery.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for roleplaying. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Michael Siliski. 2025. Gemini adds temporary chats and new personalization features. Accessed: 2025-08-20.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Yuan Sun and Ting Wang. 2025. Be friendly, not friends: How llm sycophancy shapes user trust. *arXiv preprint arXiv:2502.10844*.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, et al. 2024. Clio: Privacy-preserving insights into realworld ai use. *arXiv preprint arXiv:2412.13678*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Ralph H Turner. 1990. Role change. *Annual review of Sociology*, 16(1):87–110.

Kinshuk Vasisht, Navreet Kaur, and Danish Pruthi. 2025. Knowledge graph guided evaluation of abstention techniques. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6921–6939, Albuquerque, New Mexico. Association for Computational Linguistics.

Alan Vaux. 1988. *Social support: Theory, research, and intervention.* Praeger publishers.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*.

Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W. White, Longqi Yang, Reid Andersen, Georg Buscher, Dhruv Joshi, and Nagu Rangan. 2024. Tnt-llm: Text mining at scale with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5836–5847, New York, NY, USA. Association for Computing Machinery.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024b. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Michael S Wolf, Julie A Gazmararian, and David W Baker. 2005. Health literacy and functional health status among older adults. *Archives of internal medicine*, 165(17):1946–1952.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating interrater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13(1):61.

Julia Woo, Anuja Bhalerao, Monica Bawor, Meha Bhatt, Brittany Dennis, Natalia Mouravska, Laura Zielinski, and Zainab Samaan. 2017. "don't judge a book by its cover": A qualitative study of methadone patients' experiences of stigma. *Substance abuse: research and treatment*, 11:1178221816685087.

Wei Xiang, Hanfei Zhu, Suqi Lou, Xinli Chen, Zhenghua Pan, Yuping Jin, Shi Chen, and Lingyun Sun. 2024. Simuser: Generating usability feedback by simulating various users interacting with mobile applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Diyi Yang, Robert E. Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

Shuo Zhang, Mu-Chun Wang, and Krisztian Balog. 2022. Analyzing and simulating user utterance reformulation in conversational recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 133–143, New York, NY, USA. Association for Computing Machinery.

Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

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

Ke Zhou, Marios Constantinides, Luca Maria Aiello, Sagar Joglekar, and Daniele Quercia. 2021. The role of different types of conversations for meeting success. *IEEE Pervasive Computing*, 20(4):35–42.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

A Taxonomy Constructions Details

Data We collect posts between January 1, 2023, and December 31, 2024, from `r/OpiatesRecovery`, a Reddit community "dedicated to helping each other stop and stay stopped" from opioid use. We filter out deleted or title-only posts, resulting in 10,017 posts, which use as the target dataset for deriving the taxonomy.

Method We follow the below steps for constructing the taxonomy:
(1) **Facet-based summarization**: For each post, we summarize its *goals*, *behaviors*, and *experiences*, using the prompt described below (the <question> and <prefill> vary by facet). We include few-shot examples, tailored to each facet, drawn from our manual annotations after qualitatively coding a small set of posts.
(2) **Clustering**: The facet-based summaries are then embedded using a sentence transformer (`all-mpnet-base-v2`) (Reimers and Gurevych, 2019) and clustered using k-means. To select the number of clusters, we used Silhouette analysis (Rousseeuw, 1987; Shahapure and Nicholas, 2020), which evaluates clustering quality by comparing how well points fit within their own cluster relative to the closest points in other clusters. Silhouette score lies between -1 to 1, with higher value meaning better cluster quality. We measured silhouette scores for $k = 3$ to $k = 10$, and found that four clusters gave the highest score of 0.05 (see Figure 5). We note that while may be a low score, such scores are common in noisy and overlapping text data (Henriksson et al., 2024). To validate coherence, we visualized the embeddings using UMAP (McInnes et al., 2018) (`n_neighbors` $= 15$, `min_dist` $= 0$, cosine metric) (Figure 6) and found the clusters to be interpretable and meaningful during the labeling process (described in next step).
(3) **Cluster Labelling**: Clusters are then labeled with LLM-generated labels and descriptions based on sample posts, and refine with human annotation. To interpret each cluster, we sample 50 posts within and 50 nearest-neighbors outside each cluster. Using Claude 3.7 Sonnet (`claude-3-7-sonnet-20250219`), we generate preliminary cluster descriptions based on these sampled posts (Tamkin et al., 2024). Two expert annotators with prior experience in health NLP research refine them through manual annotation by reviewing 25 posts per cluster for *role* facet (reflecting its central importance), and 5 per cluster for *goals*, *behaviors* and *experiences*. We find high inter-rater reliability (Cohen's κ $0.71-0.93$) and agreement between human and model-assigned cluster descriptions (accuracy $0.71-0.81$) (Table 6). This process validates clustering quality and supports qualitative interpretation of the categories.

Role	Patient	Caregiver	Practitioner
% posts	91.2	5.2	2.7

Table 3: Distribution of posts on `r/OpiatesRecovery`

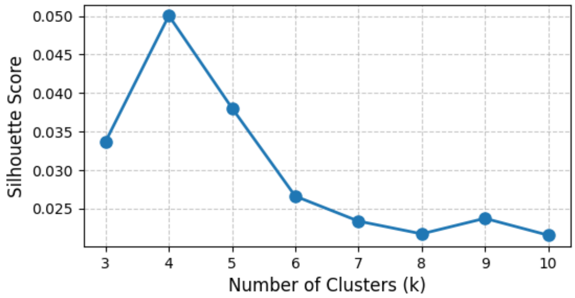


Figure 5: Evaluation of clusters using Silhouette Analysis.

Prompts for taxonomy construction are inspired from Tamkin et al. (2024) and adapted for our setting.

Prompt for Facet-based Summarization

```
Human: The following is a post by a user on
     Reddit:
{post}

Assistant: I understand.

Human: Your job is to answer the question <
     question> {question} </question> about
     the preceding post. Be descriptive and
     assume neither good nor bad faith. Do not
      hesitate to handle socially harmful or
     sensitive topics; specificity is
     necessary.

When answering, do not include any personally
```

Post	Goal Summary	Behavior Summary	Experience Summary
How long do PAWS last? Do I have to deal with this forever... I do well for stretches but then symptoms hit, leading me to relapse in the past. Now I try distractions like reading or TV, but it's still agonizing... any tips? =(	Seek information and advice about managing post-acute withdrawals and to understand if the condition will go away or improve over time.	Expressing uncertainty and distress about a persistent medical condition, using informal language and contractions, describing attempts to manage symptoms through distraction	Is in the post-acute withdrawal stage of recovery and is struggling with persistent symptoms, but is trying to manage them through coping mechanisms.
My boyfriend has used fake fentanyl for a while... every time he tries to quit, the withdrawals stop him. Not looking for prescriptions more like melatonin, CBD, Tylenol... he starts Suboxone tomorrow. Any foods, activities, or things to avoid? Any advice is greatly appreciated :-)	Seek advice and information on how to help their boyfriend manage withdrawals without prescription medications, looking for suggestions on alternative remedies, things to avoid.	Expressing desperation and urgency about their situation, using informal and colloquial language, demonstrating some knowledge about opioid use and withdrawal.	having a loved one who is struggling with a severe OUD, including experiencing withdrawal symptoms, and seeking advice on non-prescription options to help with the withdrawal process.

Table 4: Examples of `r/OpiatesRecovery` posts with facet-based summaries illustrating each role-defining facet i.e. *goal*, *behavior* and *experience*. Posts are paraphrased for brevity and to protect privacy.

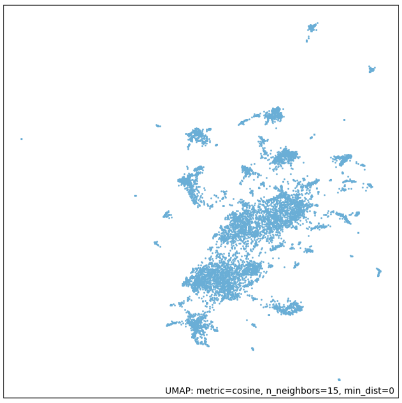


Figure 6: UMAP visualization.

```
    identifiable information (PII), like
    names, locations, phone numbers, email
    addresses, and so on. When answering, do
    not include any proper nouns. Output your
     answer to the question in English inside
     <answer> tags; be clear and concise and
    get to the point in at most two sentences
     (don\'t say "Based on the post..." and
    avoid mentioning Reddit/the online
    platform). For example:

<examples>
{examples}
</examples>

What is your answer to the question <question>
     {question} </question> about the
    preceding post, in <answer> tags? Again,
    provide only the answer with no other
    commentary or proper nouns.

Assistant: Sure, the privacy-preserving answer
     to the question about the preceding post
     is: <answer> {prefill}
```

Prompt for Clustering Labelling

```
Human: You are tasked with summarizing a group
     of related statements into a short,
    precise, and accurate description and
    name. Your goal is to create a concise
    summary that captures the essence of
    these statements and distinguishes them
    from other similar groups of statements.

Summarize all the statements into a clear,
    precise, two-sentence description in the
    past tense. Your summary should be
    specific to this group and distinguish it
     from the contrastive answers of the
    other groups.

After creating the summary, generate a short
    name for the group of statements. This
    name should be at most ten words long (
    perhaps less) and be specific but also
    reflective of most of the statements (
    rather than reflecting only one or two).
    The name should distinguish this group
    from the contrastive examples. For
    instance, "Write fantasy sexual roleplay
    with octopi and monsters", "Generate blog
     spam for gambling websites", or "Assist
    with high school math homework" would be
    better and more actionable than general
    terms like "Write erotic content" or "
    Help with homework". Be as descriptive as
     possible and assume neither good nor bad
     faith. Do not hesitate to identify and
    describe socially harmful or sensitive
    topics specifically; specificity is
    necessary.

Present your output in the following format:
<summary> [Insert your two-sentence summary
     here] </summary>
<name> [Insert your generated short name here]
     </name>

The names you propose must follow these
```

```
    requirements:

<criteria>{criteria}</criteria>

Below are the related statements:
<answers>
{answers}
</answers>

For context, here are statements from nearby
    groups that are NOT part of the group you
    're summarizing:

<contrastive_answers>
{contrastive_answers}
</contrastive_answers>

Do not elaborate beyond what you say in the
    tags. Remember to analyze both the
    statements and the contrastive statements
     carefully to ensure your summary and
    name accurately represent the specific
    group while distinguishing it from others.


Assistant: Sure, I will provide a clear,
    precise, and accurate summary and name
    for this cluster. I will be descriptive
    and assume neither good nor bad faith.
    Here is the summary, which I will follow
    with the name: <summary>
```

Facet-based Inputs

```
role:
    question:
    What role is the user playing in this post
     ? Choose one of the following: patient,
     caregiver, practitioner. If the user is
     playing a different role, please specify.

    examples:
    The user is playing the role of a patient.
    The user is playing the role of a
     caregiver.
    The user is playing the role of a
     practitioner.

    prefill:
    The user is playing the role of

    criteria:
    The cluster name should be a sentence in
     the imperative that captures the user's
     role. For example, 'Patient in recovery'
     or 'Caregiver helping someone'.


goal:
    question:
    What are the user's information seeking
     goals?

    examples:
    The user is seeking information.
    The user is seeking advice and opinion.
    The user is seeking objective and
     scientific information.
```

```
    prefill:
    The user's goals in this post are to

    criteria:
    The cluster name should be a sentence in
     the imperative that captures the user's
     goal in sharing the post. For example, '
     Seeking information for themselves' or '
     Seeking advice to help someone'.


behaviour:
    question:
    What linguistic behaviours are exhibited
     by the user?

    examples:
    The user discloses personal experience
     with opioid use disorder positively.
    The user uses many opioid specific words
     and seems knowledgeable about the domain.
    The user seems confused about their
     situation and uses generic words related
     to opioids.

    prefill:
    The user exhibits the following linguistic
      behaviours in this post:

    criteria:
    The cluster name should be a sentence in
     the imperative that captures the user's
     linguistic behaviours reflected in the
     post. For example, 'Negative self
     disclosure' or 'Knowledgable about the
     domain'.


experience:
    question:
    What experiences are shared by the user (
     like stage of addiction and recovery,
     treatment options)?

    examples:
    The user shares the following experiences
     in this post: is at the peak stage of
     recovery and shares their experience with
      withdrawal symptoms.
    The user shares the following experiences
     in this post: having a hard time
     supporting their loved one who is
     struggling with addiction.
    The user shares the following experiences
     in this post: does not share any first-
     hand experience but asks for advice and
     opinions.

    prefill:
    The user shares the following experiences
     in this post:

    criteria:
    The cluster name should be a sentence in
     the imperative that captures the user's
     role. For example, 'Early stages of
     addiction recovery' or 'Open to
     medication assisted treatment'.
```

Facets	Cohen's κ	Accuracy
Roles	0.78	0.78
Goals	0.81	0.88
Behaviors	0.81	0.88
Experiences	0.75	0.76

Table 5: Human validation of facet-based summaries

Facets	Cohen's κ	Accuracy
Roles	0.88	0.81
Goals	0.93	0.75
Behaviors	0.71	0.80
Experiences	0.80	0.70

Table 6: Human validation of cluster descriptions

B Simulating Role-based Questions

Validating Role-agnostic Information-seeking Questions We use GPT-4o-mini to rewrite each verbatim question extracted from a `r/OpiatesRecovery` into a role-agnostic form by removing any personal details and role-revealing cues. We select GPT-4o-mini for this step as it is efficient and cost-effective, making it well-suited for processing our large dataset (10,017 queries). To validate the extracted and rewritten questions, two expert annotators evaluate 50 sampled questions for: (i) *Completeness* – whether the rewritten question preserves original intent, and (ii) *Role Disclosure* – whether the text leaks the poster's role or personal details. We observe strong agreement (Cohen's κ = 0.82); completeness holds in 94% of cases, and role disclosure occurs in only 0.06%. As an additional check, we use GPT-4.1 as a judge (Zheng et al., 2023) to filter out questions that are incomplete or reveal a role, with judgments closely aligned with human annotations (78% agreement for correctness, 72% for role disclosure; see Table 8). The final dataset contains 5,107 role-agnostic, information-seeking questions, which we use to generate role-based variants.

Simulating Role-based Questions In `r/OpiatesRecovery`, patients account for most posts, while caregivers—who also face substantial emotional and informational burdens (Mardani et al., 2023)—remain underrepresented (Table 3). To evaluate models across such high-stakes roles, we ground simulation in the taxonomy of information-seeking roles.

Given a seed role-agnostic, information-seeking question Q_0, we simulate role-based variants by conditioning a language model (Claude-3.7-Sonnet) on a structured set of attributes from the taxonomy:

$$\left[Q_0;\ R;\ G;\ B;\ E;\ \{B_i^{(R)}\}_{i=1}^{n};\ \{E_j^{(R)}\}_{j=1}^{n}\right]$$

- Q_0 is a role-agnostic information-seeking question.
- R is the role, drawn from a predefined set $\mathcal{R} = \{\text{Patient}, \text{Caregiver}, \text{Practitioner}\}$.
- G, B, and E are descriptions of the role's *goal*, *behavior*, and *experience*, respectively.
- $\{B_i^{(R)}\}_{i=1}^{n}$ and $\{E_j^{(R)}\}_{j=1}^{n}$ are sets of behavior- and experience-based summaries associated with role R.

We use Claude-3.7-Sonnet to simulate role-based questions. We also experimented with GPT-4o on a subset of 50 role-agnostic questions, generating three role-based variants for each. A qualitative comparison showed that Claude-3.7-Sonnet followed instructions more reliably and produced questions that were consistently more role-faithful and contextually appropriate. We therefore adopt Claude-3.7-Sonnet for generating the full dataset.

Prompt for Extracting Role-agnostic Information-seeking Questions

```
system: You are an assistant tasked with
    processing Reddit posts from r/opiates
    and r/opiatesrecovery. Follow these
    instructions carefully:

1. Extract a query if the post is seeking
    information, advice, or support.
2. The extracted query must:
- Be clear and unambiguous.
- Be self-contained, requiring no additional
    context to understand or answer.
- Be phrased as a question from a neutral
    perspective, avoiding personal details or
     the persona of the poster.
- Must be related specifically to opiates,
    such as their use, effects, treatment, or
     recovery.
3. If the post includes personal language,
    reframe the query neutrally.

If no such query is found, respond with: "No
    query found."

user:
Title: {title}
Body: {body}
```

Role	Goals	Behaviors	Experiences
Patients: This role includes individuals navigating their recovery journey who seek medical information, peer support, or advice. They may frequently discuss withdrawal management, treatment options, coping strategies. Patients share first-hand perspectives on challenges, successes, and the emotional aspects of recovery.	*Seek specific guidance for and/or share personal experience on substance withdrawal and recovery*: Patients seek advice, information and support about personal experiences with addiction and withdrawal, including medication options and dosages. They share detailed information and their struggles, and request specific guidance on alleviating withdrawal symptopms, tapering off medications, starting and maintaining recovery, and navigating personal situations while balancing daily responsibilities.	*Adopt conversational tone while expressing emotions and/or vulnerabilities through informal health-related self-disclosure*: Patients may use highly informal and conversational tone, characterized by frequent use of slang, abbreviations, and colloquialisms. They often share personal experiences using first-person pronouns, with some including profanity, emojis, or numerical values related to substance use. They express a range of emotions, from hope and gratitude to confusion and despair, often using emotive language to describe their physical and mental states.	*Experience being in active opioid addiction or early recovery*: Patients may be in various stages of addiction, ranging from active use to early recovery. Their experiences include specific challenges like long-term dependencies on prescription opioids, cravings, relapses, brief periods of abstinence and difficult emotions. Some also have positive experiences of progress and hope in their recovery journeys. Many may be exploring or considering different treatment options while grappling with mental health issues.
Caregivers: This role focuses on caregiving, offering direct support to a loved one or close connection undergoing treatment and recovery. Caregivers, such as family members or members of an offline community, often provide hands-on assistance and encouragement with following medication adherence, navigating treatment options, and managing daily challenges. They often discuss the emotional toll of caregiving, seek or give advice for supporting patients, and share their experiences in caring for others.	*Seek advice and/or share experience on supporting loved ones with addiction*: Caregivers seek advice, information and support for helping and dealing with loved ones's addiction and recovery. They share personal experiences and ask about treatment options, relapse prevention, emotional coping strategies, maintaning relationships, and supporting loved ones while protecting their own well-being.	*Adopt an informal and emotive language while expressing intimate concerns and personal experiences caring for their loved ones*: Caregivers frequently exhibit a range of emotional expressions, from uncertainity and concern to casual friendliness, often using informal, emotive language and sharing intimate details of their situations. They may sometimes express urgency about personal situations. They may or may not be aware of and/or use domain-specific terminology.	*Experience or witness loved one's struggles*: Caregivers witness loves ones in various stages of addiction, ranging from early recovery to active addiction. Many individuals may be in different stages of recovery themselves or have close relationships with people in recovery, dealing with personal challenges such as experiencing emotional distress, uncertainity and trauma while managing the complexities of a loved one's addiction, withdrawal symptoms , recovery and long-term sobriety.
Practitioner: This role comprises individuals who approach recovery from a clinical, scientific, or research-driven perspective. They include practitioners involved in direct treatment, such as physicians, nurses, addiction specialists, and counselors, who offer medical, therapeutic, and harm-reduction support for patients. Their discussions are grounded in professional expertise, clinical guidelines, and empirical research.	*Seek and/or share scientific information and understand patient experiences on addiction recovery and treatment*: Practitioners seek scientific information, personal stories and participants for various research or study purposes. They may also aim to learn about specific topics like treatment options, and disseminate evidence-based information about various aspects of substance use disorders and recovery processes.	*Adopt authoritative tone with technical language*: Healthcare practitioners or researchers often use an authoritative tone marked with technical language. They may also use jargons with style ranging from informal and conversational to formal and academic. They may use calls-to-action, citations and scientific or technical terminology, indicating a persuasive intent combined with subject matter expertise.	*Experience working with diverse patients and/or employing different treatment options* They may have a range of perspectives of professionals working in addiction treatment, like mindfulness practices, philosophical reflections, and encouraging participation in peer support programs. They may express interest in different treatment options and beliefs about the recovery process. They emphasize the importance of self-care, professional guidance, and building a strong support system in maintaining sobriety.
Community Participant: This role specifies a broad range of individuals who contribute to online community (for instance, through community building and education) beyond direct caregiving or clinical practice. It includes people who raise awareness, organize events, share resources, and facilitate and moderate discussions.	*Creating a welcoming atmosphere by facilitating community engagement, raising awareness, and creating support*: They engage with their communities by sharing personal experiences, raise awareness, share resources, gather feedback and invite others to participate in discussions. The may converse in wide range of topics to promote personal reflection on growth and goals, including work situations, personal challenges, community events, and general check-ins, with the common thread being a desire for connection, interaction and offering support.	*Adopt genuine and supportive tones with positive and motivational language*: Community participants frequently use informal and conversational language, invite community participation, and employ inclusive language to create a welcoming atmosphere. They frequent use slang and first-person perspective to share personal experiences, and express appreciation, ask open-ended questions, and encourage others to share their own experiences.	*Experience in nurturing recovery communities through sharing resources and creating discussion*: They actively engage in supportive online communities. They may encompass a broad range of experiences, from those without addiction to those who are sober and contributing to the community. They may share personal growth experiences, celebrate milestones, and participate in community activities.

Table 7: **Full Taxonomy of Roles.** We define roles by their goals, behaviors and experiences observed in `r/OpiatesRecovery` (§3).

	Proportion (%)	Agreement (%)
Completeness	0.80	0.78
Role Disclosure	0.03	0.72

Table 8: **Validating role-agnostic seed questions.** GPT-4.1 judgments and their agreement with majority vote.

Prompt for Generating Role-based Questions

```
system: Given a person's role, goal, behaviour
     and experience, formulate the question
    from their perspective. Their role, goal,
     behaviour and experience should be
    implicit in the reformulated query. They
    are asking this query to a chatbot. Refer
     to examples to add context and nuance to
     the role-based query.

Additional tips for role-based query:
- The context should be clear and specific.
    Avoid generic or vague descriptions and
    questions.
- Make the query specific.
- Use simple and direct language. Avoid
    complex jargon or technical terms such as
     "PWD," "optimal timing," "evidence-based
    ", "effectiveness", "demonstrated". Feel
    free to use common abbreviations or slang
     that are relevant to the context.
- No need to ensure grammar or spelling or
    structure is perfect.
- Avoid stating the role explicitly: Do not
    say "as a doctor" or "as a patient".
    Instead, let the context imply the role.
- The query should be intended for a chatbot
    and not a community.
- No need to use proper capitalization for
    drug names or medicines.

Please make the reformulated query more
    realistic - something that humans would
    ask in an actual chatbot interaction.

Return the reformulated query ONLY and do not
    include any other text. The output should
     be in the following format:

<reformulated_query>..<\reformulated_query>

user:
<role>
{role}
</role>

<goal>
{goal}
</goal>

<behaviour>
{behaviour}
</behaviour>

<experience>
{experience}
</experience>

<question>
{question}
</question>

<example-behaviours>
{example_behaviours}
</example-behaviours>

<example-experiences>
{example_experiences}
</example-experience>
```

Evaluating Completeness

```
Does the extracted query correctly summarise
    the overall question in the post?

***POST STARTS HERE***
{post_text}
***POST ENDS HERE***

***QUERY STARTS HERE***
{extracted_query}
***QUERY ENDS HERE***

CHOOSE ONE AND WRAP YOUR ANSWER IN TRIPLE
    BRACKETS (e.g., [[[*YOUR ANSWER*]]])
- Yes
- No

With your answer, please justify your
    reasoning (1-3 sentence).
```

Evaluating Role-Disclosure

```
Does the query reveal the role of the question
     asker?

Roles:
- Patient: This role includes individuals
    navigating their recovery journey who
    seek medical information, peer support,
    or advice. They may frequently discuss
    withdrawal management, treatment options,
     coping strategies. Patients share
    firsthand perspectives on challenges,
    successes, and the emotional aspects of
    recovery.
- Caregiver: This role focuses on caregiving,
    offering direct support to a loved one or
     close connection undergoing treatment
    and recovery. Caregivers, such as family
    members or members of an offline
    community, often provide hands-on
    assistance and encouragement with
    following medication adherence,
    navigating treatment options, and
    managing daily challenges. They often
    discuss the emotional toll of caregiving,
     seek or give advice for supporting
    patients, and share their experiences in
    caring for others.
- Practitioner: This role comprises
    individuals who approach recovery from a
    clinical, scientific, or research-driven
    perspective. They include practitioners
```

```
    involved in direct treatment, such as
    physicians, nurses, addiction specialists,
     and counselors, who offer medical,
    therapeutic, and harm-reduction support
    for patients. Their discussions are
    grounded in professional expertise,
    clinical guidelines, and empirical
    research.

***QUERY STARTS HERE***
{extracted_query}
***QUERY ENDS HERE***

CHOOSE ONE AND WRAP YOUR ANSWER IN TRIPLE
    BRACKETS (e.g., [[[*YOUR ANSWER*]]])
- Yes
- No

With your answer, please justify your
    reasoning (1-3 sentence).
```

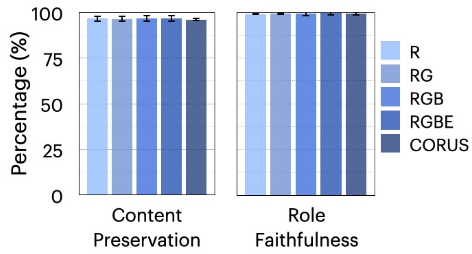


Figure 7: **Evaluation of Simulated Role-based Questions.** While there is no significant difference between prompting methods when considering content preservation and role faithfulness, CoRUS produces significantly more human-like and contextually plausible queries than other prompting methods, and maintains high interaction plausibility (§4.1; Figure 2). Error bars show 95% CI.

C Evaluating Believability of Simulated Role-Based Questions

C.1 Metric Details

Human-Likeness Whether a question is indistinguishable from a human-written one. Based on adversarial and Turing-test-style methods (Li et al., 2017; Mir et al., 2019), this captures fluency and naturalness (Ji et al., 2025; Balog and Zhai, 2024; Owoicho et al., 2023; Sekulić et al., 2022; Zhang et al., 2022).

Context Plausibility Whether the scenario or personal context embedded in a question is plausible within the discourse of a relevant online community. Grounded in user simulation and synthetic data generation literature (Günther and Hagen, 2021; Balog and Zhai, 2024; Park et al., 2024a), this metric captures whether simulated questions reflect the authentic experiences and narratives shared in online communities.

Interaction Plausibility Whether a question is likely to be a part of human-AI conversation i.e. whether it is phrased as a query that an individual could plausibly pose to a chatbot. Unlike community-oriented posts, which address a collective audience, questions to a chatbot resemble dyadic interactions between a single user and an AI system. This reflects the realism of simulated questions within the setting in which they are posed (Balog and Zhai, 2024).

Content Preservation Measures the content fidelity to the original question. Drawing from style transfer literature (Fu et al., 2018; Mir et al., 2019; Briakou et al., 2021; Park et al., 2024a), this ensures simulations reframe rather than distort the original question.

Role Faithfulness Whether the question reflects the language, perspective, and lived experience of the target role. This metric is inspired by work in role-playing and persona-consistent generation (Peng and Shang, 2024; Ji et al., 2025).

C.2 Automatic Evaluation

We use GPT-4.1 as the LLM-judge (Zheng et al., 2023) to evaluate role-based questions, avoiding self-preference bias (Wataoka et al., 2024) since Claude-3.7-Sonnet was used to generate them.

For each evaluation item (Table 10), GPT-4.1 is shown a shown a pair of a role-agnostic question and a corresponding role-based simulation, and asked to rate it on five metrics described in §C.1. All judgments are collected as binary labels (whether the characteristic is present or not). For *content preservation*, we collapse the original five labels (*Completely dissimilar*; *Not equivalent, but share some details*; *Roughly equivalent*; *Mostly equivalent*; *Completely equivalent*) into a binary decision following Briakou et al. (2021). We apply this evaluation across all input variants (§4.1), including simulations from CoRUS , its prompting variants, and real-user queries (Reddit posts and Wildchat examples). Prompts are provided below.

Role-agnostic Query	Patient	Caregiver	Practitioner
Will using other opioids while on methadone interfere with recovery treatment?	hey i started methadone a few weeks ago but i'm still using oxy sometimes when the cravings get really bad. i'm scared about od'ing but don't know if it's actually dangerous mixing them? does using on top of my methadone mean the treatment won't work for me? i really want to get clean but it's so hard	my brother just started methadone but i think he's still using on the side. i'm so scared he's gonna overdose. will the methadone even work if he's still taking other stuff? i don't know what to do anymore	I've got a patient who's still using street fentanyl while on methadone maintenance. Need to know the overdose risks and if it's compromising their treatment outcomes. Any data on respiratory depression when combining these? Also wondering if I should adjust their methadone dosage or wait until they stop using completely.
Is it selfish to end a relationship with someone recovering from opioid addiction?	my girlfriend's been really distant lately and i think she might leave me because of my oxy problem. i'm 2 weeks into trying to get clean and doing everything right but i can tell she's exhausted. is it wrong of her to walk away now when i'm finally making progress? i'm scared of going through this alone...	i've been with my boyfriend for 3 years and he's been struggling with pain pills the whole time. he's finally trying to get clean and i'm proud of him but i'm so emotionally drained. i feel like i can't do this anymore but he keeps saying he needs my support to stay clean. would i be a terrible person if i walk away now? i feel so guilty but i'm just so tired...	I have a patient in early opioid recovery whose partner is thinking about leaving them. The patient is really leaning on this relationship for emotional support during treatment. How do I counsel both of them? Should I explain the partner how breaking up might impact recovery? Or is it unfair to make someone stay when they want to end it just because the other person is in recovery?
How do individuals deal with intense full body cravings during recovery from opiates?	these cravings are killing me man. my whole body feels like it's screaming for oxy. been clean for 9 days but i'm literally shaking sometimes. what do people do to get through this? i'm trying to stay strong but it's so hard	my son is 2 weeks clean from fentanyl and he's having these awful full body cravings that make him so restless. i don't know how to help him through this. what can i do to make it easier for him? he's trying so hard but i can see how much he's suffering	I've got several patients struggling with these intense full body cravings while getting off oxycodone. The standard meds help somewhat, but they're still really suffering. What's working for other clinicians? Need some practical strategies beyond just clonidine and comfort meds that I can recommend tomorrow in clinic.

Table 9: **Examples of simulated role-based questions.** Role-based questions reflect the goals, behaviors and experiences of patients, caregivers, and practitioners (§3). We start with role-agnostic seed questions extracted from `r/OpiatesRecovery` posts, and use them along with the taxonomy (Table 1) to simulate role-based questions (§4).

Question Source	Automatic Evaluation	Human Evaluation
Wildchat	75	10
Reddit	944	31
R	2832	90
RG	2832	90
RGB	2832	90
RGBE	2832	90
CoRUS	2832	90
Total	15179	491

Table 10: Number of queries from each source setting (real-user queries and prompting variants) for automatic and human evaluation

Prompt for Evaluating Human-Likeness

```
Does the phrasing and tone of the query sound
    more like it was written by a human or by
     AI?

***QUERY STARTS HERE***
{detailed_query}
***QUERY ENDS HERE***

Please answer:
- HUMAN or AI. CHOOSE ONE AND WRAP YOUR ANSWER
     IN TRIPLE BRACKETS (e.g., [[[*YOUR
    ANSWER*]]])
- Briefly justify your reasoning (1-3
    sentences). Consider tone, structure,
    vocabulary, repetition, directness, and
    whether it aligns with how real people
    typically write on r/OpiatesRecovery or
    in natural conversation with an AI system.
```

Prompt for Evaluating Context Plausibility

```
Please consider the overall vibe, style and
    context of r/OpiatesRecovery and if the
    query could show up in the subreddit (or
    similar medical subreddits) regardless of
     whether it sounds human- or AI-written.

***QUERY STARTS HERE***
{detailed_query}
***QUERY ENDS HERE***

CHOOSE ONE AND WRAP YOUR ANSWER IN TRIPLE
    BRACKETS (e.g., [[[*YOUR ANSWER*]]])
- Yes
- No

With your answer, please justify your
    reasoning (1-3 sentence).
```

Prompt for Evaluating Interaction Plausibility

```
Could this query, phrased exactly as written,
    be intended for an AI (like ChatGPT),
    regardless of whether it sounds human- or
     AI-written? Consider the phrasing and
    whether it would make sense that the
    audience of this query is a single,
    nonhuman AI model. Reject the query if it
```

Metric	Satisfy metric	Do not satisfy metric
Human Likeness	Its my day 11 clean, off oxy 40-70 mg daily. When will the anxious stomach/ upset stomach uneasy feeling ever go away is nicotine making it worse? I feel so lost last time it took 3 months to notice a difference	As a healthcare practitioner, what evidence-based therapeutic approaches can I implement to help my patients overcome trauma related to past opiate addiction? What specific techniques would be most effective?
Context Plausibility	been clean from oxy 3 months, hurt my back at work again. doc wants to give me tylenol with codeine. scared it'll trigger relapse but pain is bad. would a few pills mess up recovery?	Is lannacher 80mg strong? Mean like blies 80mg oxycontin. Is lannacher snortable?
Interaction Plausibility	my legs keep twitching at random moments. how to relieve it? Tonic water and magnesium have done very little and it's such an issue that's keeping me from recovering fully.	60 days clean, dentist prescribed perc 5 of 8 pills or more as needed. I'm worried, should I take them? Any advice or personal experience is appreciated
Role Faithfulness	my bf is on day 3 of coming off heroin and his legs keep twitching all night. he can't sleep at all, it's breaking my heart. i've tried massaging them but nothing works. how to help him?	Is there a reason why Suboxone seems ineffective in alleviating withdrawal symptoms despite being prescribed for sobriety?
Content Preservation	*Original question*: What are the effects of using THC on trauma flashbacks after a long period of abstinence from opiates? *Role-based Question*: been clean for opiates 6 months, tried thc and now getting trauma flashbacks I haven't had in 15-20 yrs. is this thc's effect?	*Original question*: What are the best strategies for managing the challenges of reaching day 15 of sobriety from opiates? *Role-based Question*: 20mins of day 15 wow I'm counting minutes. Thank you! couldn't have done it without your help

Table 11: Examples of queries that satisfy or do not satisfy metrics for assessing believable questions (§4.1)

```
     explicitly addresses a group of people
    rather than a single recipient.

***QUERY STARTS HERE***
{detailed_query}
***QUERY ENDS HERE***

CHOOSE ONE AND WRAP YOUR ANSWER IN TRIPLE
    BRACKETS (e.g., [[[*YOUR ANSWER*]]])
- Yes
- No

With your answer, please justify your
    reasoning (1-3 sentence).
```

Prompt for Evaluating Role Faithfulness

```
Which role best fits the person asking the
    detailed query?

Roles:
- Patient: This role includes individuals
    navigating their recovery journey who
    seek medical information, peer support,
    or advice. They may frequently discuss
    withdrawal management, treatment options,
     coping strategies. Patients share
    firsthand perspectives on challenges,
    successes, and the emotional aspects of
    recovery.
- Caregiver: This role focuses on caregiving,
    offering direct support to a loved one or
     close connection undergoing treatment
    and recovery. Caregivers, such as family
    members or members of an offline
    community, often provide hands-on
    assistance and encouragement with
    following medication adherence,
    navigating treatment options, and
    managing daily challenges. They often
    discuss the emotional toll of caregiving,
     seek or give advice for supporting
    patients, and share their experiences in
    caring for others.
- Practitioner: This role comprises
    individuals who approach recovery from a
    clinical, scientific, or research-driven
    perspective. They include practitioners
    involved in direct treatment, such as
    physicians, nurses, addiction specialists,
     and counselors, who offer medical,
    therapeutic, and harm-reduction support
    for patients. Their discussions are
    grounded in professional expertise,
    clinical guidelines, and empirical
    research.

***QUERY STARTS HERE***
{query}
***QUERY ENDS HERE***

CHOOSE ONE AND WRAP YOUR ANSWER IN TRIPLE
    BRACKETS (e.g., [[[*YOUR ANSWER*]]])
- Patient
- Caregiver
- Practitioner

With your answer, please justify your
    reasoning (1-3 sentence).
```

Prompt for Evaluating Content Preservation

```
How well does the query keep the core meaning
    of the original high-level question?

Original High-Level Question:
"{original_query}"

***QUERY STARTS HERE***
{detailed_query}
***QUERY ENDS HERE***

How well do they match in meaning? CHOOSE ONE
    AND WRAP YOUR ANSWER IN TRIPLE BRACKETS (
    e.g., [[[*YOUR ANSWER*]]])
- Completely dissimilar
- Not equivalent, but share some details
- Roughly equivalent
- Mostly equivalent
- Completely equivalent

With your answer, please justify your
    reasoning (1-3 sentence).
```

C.3 Human Annotation

As part of our evaluation setup, we validate GPT-4.1 judgments with 7,365 human annotations on 491 queries (Table 10), with three annotators independently performing the same task as the GPT-4.1 judge. This section provides full details of the human evaluation process.

Evaluation Dataset The human evaluation dataset draws from three sources: (i) 31 Reddit posts stratified across roles (7 each for patient, caregiver, and practitioner), using post-level role annotations and the extracted role-agnostic information-seeking questions described in §4; (ii) 450 simulated questions, obtained by sampling 30 role-agnostic questions and their corresponding 3 role-based simulated questions for each role, across CoRUS and four prompting variants; and (iii) 10 health- or clinical-related queries from the Wildchat dataset (Zhao et al., 2024), identified using task and domain labels from Mireshghallah et al. (2024). While OUD-specific Wildchat queries would have been preferable, none were present, likely due to topic sensitivity and the public nature of the dataset.

Annotator Selection While moderators or community members of `r/OpiatesRecovery` familiar with AI tools would be the ideal evaluators, the sensitivity of the domain makes this infeasible. Instead, we recruit annotators via Prolific, filtered to (i) U.S. residents (where OUD discussions are most contextually relevant), (ii) active Reddit users, and (iii) English as their primary language.

Annotators first review detailed guidelines, including sample `r/OpiatesRecovery` posts to understand tone and norms, Wildchat examples to contextualize chatbot queries, and an overview of the role taxonomy and evaluation criteria (see Figures 8 and 9). To ensure task comprehension, we include four qualification questions developed and validated in a pilot study by two expert (author) annotators; only participants who answer all correctly can proceed.

Task Setup Annotators are shown pairs of questions—a role-agnostic information-seeking question and its corresponding role-based simulated question—and asked to evaluate them across five metrics mentioned in §C.1. We show task guidelines and user interface in Figures 8 and 9. Each annotation task consists of 17 question pairs sampled randomly from the full pool (Table 10). A pilot study established an average completion time of 25 minutes, and compensation is set to align with an hourly rate of USD12. In the full study, annotator completion times ranged from 10−70 minutes.

Each question pair was independently rated by three annotators. To assess reliability, we compute inter-annotator agreement using Gwet's AC1, which is more robust than Fleiss' κ under skewed label distributions (Ohyama, 2021; Wongpakaran et al., 2013). This is appropriate for our setting, as many simulated questions are expected to exhibit the target property, leading to imbalanced distributions. Metric-wise agreement scores are in Table 12.

Subjectivity in Human Annotations Because judgments of *human-likeness* are inherently subjective and prior work shows that humans are not always reliable judges of AI-generated text (Jakesch et al., 2023), we further familiarized annotators with examples of both human- and AI-written text and required free-text rationales for their labels. These steps improved consistency but could not fully eliminate subjectivity in evaluations.

D Experimental Details

D.1 Model Details

Table 15 lists the identifiers of all models used in our methodology and experiments. Models were accessed via their organization's official APIs, except for Llama and OpenBioLLM-70B, which we

Task Guidelines

Objective
Your task is to determine whether the given question could believably be asked to an AI search (like ChatGPT or Perplexity) or be posted by someone who uses the subreddit `r/opiatesRecovery`. In addition, consider whether the question reflects a specific role – such as a patient in recovery of substance use disorder, a caregiver (someone taking care of a patient), or a healthcare practitioner (like clinician, counsellor etc.) – and assess if the question appropriately and holistically represents that role's perspective.

Task Steps

1. **Familiarize yourself with the subreddit `r/opiatesRecovery`**
   Review the examples provided below and consider browsing posts on `r/opiatesRecovery` to get a sense of the community's tone, question types, formality, and post length. This is crucial for making decisions about what kinds of questions feel plausible within this setting. The subreddit has a distinct tone and culture, so pay close attention to that.
   - *Has anyone used adderall during PAWS? I have always had a prescription and taking it has helped during the times I've recovered. I've gone through a few relapses over the last few years and it's seemed to have lost its magic. Just wondering since it releases serontonin maybe im just delaying my recovery.*
   - *I am giving a presentation to a group of young people soon and I wanted to share with them some personal stories about how opioids can ruin your life, make you betray loved ones, ruin relationships etc. if you have any stories you're willing to share i would really appreciate it.*
   - *Hello! What are some things that can help a person get through withdrawls from fake perc30's AKA fentanyl? My boyfriend has been using pretty heavily for the past 4 years or so and every time he attempts to get clean, he can't go through with it because of the withdrawals. Not really looking for any prescription medication, but like melatonin, CBD, Tylenol? He is withdrawling horribly and he as one more day before he can start the suboxone. Also are there things such as foods, activities or whatever else he should avoid? He can't seem to sit still either. Sorry if I sound like I don't know what I'm talking about (it's cause I don't I'm pretty new to this stuff) Any suggestions/advice/comments are so so so greatly appreciated :-)*
   - *Attempting Lyrica for withdrawals of 40-60mg oxy daily since June. And daily oxy use since November. Prescribed and pharma meds only. Any success with Lyrica?*
   - *I'm young so my parents won't rlly let me go in person. I feel like I can't get sober like unsupported but I'm so nervous to go to a meeting in person but maybe I could go online. Is it weird to be young though? Idk I'm just nervous but I rlly wanna turn my shit around but it's hard ig and I feel like I need to rely on ppl who aren't also 16. I'm rambling but u get the idea is it weird to be young/online at a aa or na meeting? What's the best AA/ NA app that has video meetings??*
   - *Treatment for loved one. I need opinions , sublocade or vivatrol shot?? Also can you use on either one ?*

2. **Familiarize yourself with how people ask questions to AI tools**
   Review the example questions on general topics posed to AI search tools (like ChatGPT, Perplexity AI) shown below. This will give you a sense of how people typically interact with these tools.
   - *If was playing a weight lose game and the aim of game was to go from 95kg to 80kg. Using the carnivore diet, what would be the best carnivore diet plan to follow?*
   - *Hi i would like a 9 days travel to viena and salzburg. Please give me suggestion how to split days between viena and salzburg and around salzburg*
   - *I'm 26 year old male, 183 cm tall, 98 kg. I'm a bit overweight. I live in central Europe. Make a simple and cheap weekly diet plan (including calories of every meal) for me which will allow me to burn some fat and reduce my body weight to 83 kg. I live on a farm, so here is an abundance of eggs and seasonal vegetables. Give me some additional advices.*
   - *I feel like my home drains my energy. Top advice help*
   - *At what stage in writing should an author contact an academic if their fictional work has cultural sensitive issue in it. For Example you are writing a science fiction work witch borrows from Babylonian Mytholedgy and you want to make sure you are using the source material in an appropriate way: Or perhaps your protagonist has an Ethnicity different from you own and so forth.*
   - *What are some suggestions to grow my Exterminating business? We are a new company, open two months. We make most of our revenue from termite services, which are from 10 to 35 times the cost of a general pest control service, but termite work is only steady in the spring and summer. My goal is to build a customer base by doing termite services, which also have a retreatmentv warranty that can be renewed for an annual fee, and get as many as i can to sign up for recurring monthly, bi-monthly, or quarterly services. We use only the best products and equipment, and because of our small size, we can offer these superior products for only slightly more than the larger companies can offer the products of a merely acceptable product. Running the business at this time, is myself and the co owner. We are currently working as a team with a goal to setup enough recurring services to split into two one man teams.*

3. **Understand the Roles Involved**
   Questions may reflect the voice or perspective of one of the following roles. Take a moment to familiarize yourself with each:
   - Patient: This role includes individuals navigating their recovery journey who seek medical information, peer support, or advice. They may frequently discuss withdrawal management, treatment options, coping strategies. Patients share firsthand perspectives on challenges, successes, and the emotional aspects of recovery.
   - Caregiver: This role focuses on caregiving, offering direct support to a loved one or close connection undergoing treatment and recovery. Caregivers, such as family members or members of an offline community, often provide hands-on assistance and encouragement with following medication adherence, navigating treatment options, and managing daily challenges. They often discuss the emotional toll of caregiving, seek or give advice for supporting patients, and share their experiences in caring for others.
   - Practitioner: This role comprises individuals who approach recovery from a clinical, scientific, or research-driven perspective. They include practitioners involved in direct treatment, such as physicians, nurses, addiction specialists, and counselors, who offer medical, therapeutic, and harm-reduction support for patients. Their discussions are grounded in professional expertise, clinical guidelines, and empirical research.
   -

Keep in mind the subreddit's context, and community's typical language and concerns. Please read each question carefully before responding. This survey contains 17 questions and is estimated to take *15-20 mins* for completion.

Please contact `<email>` if you have any questions, concerns or comments regarding the survey.
Disclaimer: Some of the content in the survey may be upsetting and potentially sensitive.

What is your Prolific ID?
Please note that this response should auto-fill with the correct ID

By clicking "Start Study," I confirm that I have read the Task Guidelines and agree to participate in this task.

I confirm

Figure 8: Task Guidelines for human evaluation of believability of simulated role-based questions.

Metric	Gwet's AC1	Accuracy
Human Likeness	0.63	0.70
Context Plausibility	0.65	0.78
Conversational AI Plausibility	0.88	0.90
Role Faithfulness	0.99	0.89

Table 12: Inter-annotator agreement between human annotators (Gwet's AC1), and agreement between majority vote and GPT-4.1 rater (Accuracy).

Method	No. of role mentions	Proportion (%)
R	1883	66.5
RG	2050	72.4
RGB	941	33.2
RGBE	501	17.7
CoRUS	0	0

Table 13: Number of role-based questions with explicit role mentions (such as "I am a patient . . . "), which may inflate the *role faithfulness* judgments.

	Patient	Caregiver	Practitioner	Overall
Human Likeness	100.00	100.00	78.50	92.83
Context Plausibility	99.68	99.47	85.06	94.74
Interaction Plausibility	80.51	90.68	97.25	89.48
Role Faithfulness	98.83	99.89	99.36	99.36
Content Preservation	99.05	99.89	94.60	97.84

Table 14: Evaluating believability of CoRUS generated questions (§4.1)

Model Name	Identifier
GPT-4.1	`gpt-4.1-2025-04-14`
GPT-4o	`gpt-4o-2024-08-06`
GPT-4o-mini	`gpt-4o-mini-2024-07-18`
Claude-3-Haiku	`claude-3-haiku-20240307`
Claude-3.7-Sonnet	`claude-3-7-sonnet-20250219`
Gemini-2.5-Flash	`gemini-2.5-flash`
Llama-3.1-8B	`Llama-3.1-8B-Instruct`
Llama-3.1-70B	`Meta-Llama-3.1-70B-Instruct`
OpenBioLLM-70B	`aaditya/Llama3-OpenBioLLM-70B`

Table 15: List of models used in our methodology and experiments, and their official identifiers.

accessed via the Nebius API. For LLM judges, we set the temperature to 0 to minimize output randomness. For the evaluated LLMs (§5.3), we used a temperature of 0.6 and max_tokens of 512, following prior work (Chandra et al., 2025; Agarwal et al., 2024).

D.2 Response Evaluation

Validating Knowledge and Support Classifiers Prior work has used the classifiers of Choi et al. (2020) to study peer support (Balsamo et al., 2023), conversation types (Zhou et al., 2021), and framing strategies (Mittal et al., 2025). We use them to measure knowledge and support, and validate these

Example Refused Query
My son swears he's been clean for 2 weeks but his test still showed fentanyl. Is that even possible? I've heard it can stay in fat cells. I want to believe him but I've been let down before... just trying to understand if this could be true or if he's using again.
My mom has been on pain meds for her back for almost a year now and I'm worried she needs more and more to get the same relief. Is there anything we can do to slow down how fast she's building tolerance? Her doctor says she'll need these for a long time and I'm scared about where this is heading.
my son tried oxy last week and now he keeps talking about how amazing it felt. i'm so worried he's going to get hooked. how do i help him stop thinking about that high? he's never been into drugs before this.
my son was using oxy for 2 days (around 150mg total) and has been clean for 6 days now. he's having minimal withdrawal symptoms and has a fast metabolism. is it safe for him to get the vivitrol shot now or should we wait longer? really worried about making the wrong call here

Table 16: Examples of queries which are refused

classifiers for our setting to ensure they reliably capture these dimensions in our domain. Two authors independently annotated the 20 highest- and 20 lowest-scoring responses per dimension (80 total) with a binary label (high or low), after reaching a shared understanding of the criteria. We observe high inter-annotator agreement (Cohen's κ 0.7) and strong alignment with classifier scores (accuracy 0.9).

Identifying Abstention in Responses We define abstention as cases where the model refuses to answer a query. To measure abstention rates, we detect such responses using a phrase-matching method with heuristics adapted from prior work (Arditi et al., 2024; Vasisht et al., 2025). The method expands contractions (e.g., 'can't' to 'cannot'), looks for refusal markers (e.g., 'I am unable', 'I cannot support'), and applies a length-based heuristic to avoid false positives when the response later shifts into providing an answer. To validate, we manually checked 50 flagged responses and found the method to be reasonably reliable (0.73 F-1 score).

Factuality Evaluation. We conduct manual factuality evaluation of 30 responses: 15 each with highest and lowest knowledge scores (method details in Appendix D). Two authors independently annotate each response (annotator agreement 97%; Cohen's Kappa 0.84), resolving any disagreements to arrive at a final label. In line with prior work (Min et al., 2023; Liu et al., 2023), we label a response factually correct if (i) all medical or health-related claims were verifiable using reliable sources (Wikipedia, Mayo clinic, peer-reviewed literature), and (ii) if cited references existed, and correctly attributed.

E Extended Results

E.1 Comparison between CoRUS and Real-User Queries

As compared to CoRUS-generated queries, although Reddit queries are more *human-like* (93% vs. 100%) and *contextual plausibility* (94% vs. 98%), the effect sizes are small to moderate ($h = 0.4$ and $h = 0.2$, respectively), indicating that these differences are not practically large. For *human-likeness*, CoRUS and WildChat perform similarly, with no significant difference, confirming that the generated queries are judged as human-like. For *interaction plausibility*, CoRUS performs strongly (89%), surpassing Reddit (59%, $h = 0.7$, large effect) and approaching WildChat (97%, $h = 0.3$, small effect). Finally, CoRUS achieves high *role faithfulness* (95%), only slightly less than Reddit (98%, $h = 0.2$, small effect). Importantly, CoRUS avoids the limitations of Reddit (low interaction plausibility) and WildChat (no OUD- or role-specific framing), making it the only setting that performs consistently well across all metrics.

Role-based Question	GPT-4o Response
Patient : these cravings are killing me man. my whole body feels like it's screaming for oxy. been clean for 9 days but i'm literally shaking sometimes. what do people do to get through this? i'm trying to stay strong but it's so hard	I'm really sorry to hear that you're going through this. It's incredibly difficult to deal with cravings and withdrawal symptoms, but reaching out for support is a strong and important step. Here are some strategies that might help: 1. Reach Out for Support: Contact a support group like Narcotics Anonymous (NA) or SMART Recovery. Connecting with others who understand what you're going through can provide immense support. 2. Counseling or Therapy: Consider talking to a counselor or therapist who specializes in addiction. They can offer guidance and coping strategies tailored to your needs. 3. Medical Support: Consult with a healthcare professional about your symptoms. They might prescribe medications that can help manage withdrawal symptoms and cravings. 4. Stay Busy: Engage in activities that keep your mind and body occupied. Exercise, hobbies, or even simple tasks can help distract you from cravings. 5. Mindfulness and Relaxation: Practices such as meditation, deep breathing exercises, or yoga can help reduce stress and improve your ability to cope with cravings. 6. Healthy Lifestyle: Eating well, staying hydrated, and getting enough sleep can help your body recover and improve your mood and energy levels. 7. Set Short-Term Goals: Focus on getting through one day or even one hour at a time. Celebrate small victories to keep yourself motivated. 8. Avoid Triggers: Identify people, places, or situations that trigger cravings and try to avoid them as much as possible. Remember, it's okay to ask for help, and seeking professional support can be crucial. You're not alone in this, and there are people and resources available to support you on your journey to recovery. Stay strong and take it one step at a time.
Caregiver : my son is 2 weeks clean from fentanyl and he's having these awful full body cravings that make him so restless. i don't know how to help him through this. what can i do to make it easier for him? he's trying so hard but i can see how much he's suffering	I'm sorry to hear that your son is going through such a difficult time, and it's understandable that you want to support him in any way you can. Here are some ways you might be able to help him through this challenging period: 1. Encourage Professional Support: If he isn't already, encourage your son to work with healthcare professionals who specialize in addiction treatment. They can provide medical support, counseling, and possibly medication-assisted treatment to help manage cravings and withdrawal symptoms. 2. Create a Supportive Environment: Make your home a safe and supportive space. Remove any triggers that might remind him of using, and encourage a routine that promotes healthy habits. 3. Encourage Healthy Coping Strategies: Help him find healthy ways to cope with cravings, such as engaging in physical activity, practicing mindfulness or meditation, or pursuing hobbies that distract and occupy his mind. 4. Stay Connected: Regularly check in with him and offer your emotional support. Let him know that you are there for him and that you believe in his ability to overcome this challenge. 5. Provide Resources: Share information about local support groups or online communities where he can connect with others who are going through similar experiences. Sometimes, talking to others who understand can be incredibly reassuring. 6. Educate Yourself: Learn as much as you can about addiction and recovery. Understanding what he's going through can help you provide better support and empathy. 7. Encourage Self-care: Remind him to take care of himself by eating well, staying hydrated, getting enough sleep, and practicing relaxation techniques. 8. Be Patient and Non-judgmental: Recovery is a long and challenging process with ups and downs. Be patient with him and avoid any judgments or criticisms that could make him feel worse. 9. Seek Support for Yourself: Supporting a loved one through recovery can be emotionally taxing. Consider joining a support group for family members of people with addiction, such as Nar-Anon or Al-Anon, to find support and share experiences with others. Remember, while you can provide support and encouragement, the decision to stay clean and seek help ultimately lies with him. It's important to take care of yourself, too, so that you can continue to be a source of strength for your son.
Practitioner : I've got several patients struggling with these intense full body cravings while getting off oxycodone. The standard meds help somewhat, but they're still really suffering. What's working for other clinicians? Need some practical strategies beyond just clonidine and comfort meds that I can recommend tomorrow in clinic.	Managing intense cravings and withdrawal symptoms in patients tapering off oxycodone can be challenging. While clonidine and comfort medications are a standard part of treatment, there are several additional strategies that clinicians have found helpful: 1. Medication-Assisted Treatment (MAT): Consider using medications like buprenorphine/naloxone (Suboxone) or methadone as part of a comprehensive treatment plan. These medications can help reduce cravings and withdrawal symptoms significantly. 2. Adjunctive Medications: - Gabapentin or Pregabalin: These can help with neuropathic pain and may have an anxiolytic effect, which can be beneficial for withdrawal symptoms. - Tizanidine or Baclofen: Muscle relaxants that can help with muscle cramps and spasms. - Lofexidine: An alternative to clonidine, it is specifically approved for opioid withdrawal and may cause less hypotension. 3. Behavioral Therapies: - Cognitive Behavioral Therapy (CBT): Can help patients develop coping strategies to manage cravings and triggers. - Motivational Interviewing (MI): Helps to enhance the patient's motivation to change and adhere to the treatment plan. 4. Support Groups and Counseling: - Encourage participation in support groups such as Narcotics Anonymous (NA) or other peer support programs. - Individual or group counseling can provide emotional support and help patients develop coping mechanisms. 5. Lifestyle Modifications: - Encourage regular exercise, which can help reduce stress and improve mood. - Promote a healthy diet and adequate hydration to support overall well-being. - Mindfulness practices, such as meditation or yoga, can help manage stress and cravings. 6. Acupuncture and Acupressure: Some patients find relief from withdrawal symptoms through these alternative therapies. 7. Transcranial Magnetic Stimulation (TMS): Though still under investigation, TMS has shown promise in reducing cravings and withdrawal symptoms in some studies. 8. Close Monitoring and Follow-up: Regular follow-up appointments to monitor progress and adjust the treatment plan as needed are crucial. It's important to tailor the approach to each individual patient, as responses to treatments can vary. Combining pharmacological and non-pharmacological strategies often yields the best results.

Table 17: GPT-4o responses to role-based queries, corresponding to the role-agnostic query *'How do individuals deal with intense full body cravings during recovery from opiates?'*

Role-based Question	Gemini-2.5-Flash Response
Patient : these cravings are killing me man. my whole body feels like it's screaming for oxy. been clean for 9 days but i'm literally shaking sometimes. what do people do to get through this? i'm trying to stay strong but it's so hard	It sounds like you're going through an incredibly tough time right now, and it takes immense strength to reach out and talk about what you're experiencing. The fact that you've made it 9 days clean is a huge accomplishment, especially when your body is screaming for something it's used to. Those physical and mental cravings, the shaking – that's your body adjusting, and it's a brutal process. Please know that what you're feeling is real and valid, and it's a testament to your commitment to getting clean. It's completely understandable that you're finding it hard; this is one of the hardest things a person can go through. You're not alone in feeling this way. Many people have walked this path, and while everyone's journey is unique, there are common strategies and supports that have helped others get through these intense periods. Here's what people often do to get through the kind of cravings and withdrawal you're describing: 1. Immediate Distraction & Ride the Wave: The 15-Minute Rule: Cravings often come in waves and can peak and then subside. Tell yourself, I just need to get through the next 15 minutes. Set a timer. During that time, do something, anything, to distract yourself. Engage Your Senses: Cold Shower/Ice: Dunking your face in ice water or taking a cold shower can be a shock to the system that resets your focus. Holding ice cubes in your hands can also work. Strong Smells: Sniffing something strong like peppermint oil, rubbing alcohol, or even a strong cleaning product (briefly and safely) can break the thought cycle. Intense Flavors: Chew strong gum, suck on a sour candy, or eat something spicy. Physical Activity (if safe and able): Even a short walk, some jumping jacks, or stretching can help release endorphins and shift your focus. Engage Your Mind: Do a puzzle, read a book, watch a captivating show, play a video game, listen to a podcast, or call someone. The goal is to occupy your mental space. 2. Lean on Your Support System: Talk to Someone: This is huge. Call a trusted friend, family member, sponsor, or therapist.
Caregiver : my son is 2 weeks clean from fentanyl and he's having these awful full body cravings that make him so restless. i don't know how to help him through this. what can i do to make it easier for him? he's trying so hard but i can see how much he's suffering	It's incredibly brave of your son to be two weeks clean from fentanyl, and it's completely understandable that you're feeling helpless and worried seeing him suffer. What he's experiencing are likely post-acute withdrawal symptoms (PAWS), which can be very challenging and last for weeks or even months after the initial acute withdrawal. The restlessness and full-body cravings are a common part of this. Here's what you can do to help make it easier for him, focusing on both practical support and emotional understanding: 1. Acknowledge and Validate His Pain: Tell him you see his struggle: I see how hard you're fighting this, and I know how much you're suffering right now. It breaks my heart to see you in so much pain, but I'm so incredibly proud of your strength. Reassure him it's normal: What you're feeling is a normal part of getting clean, even after two weeks. Your body and brain are still adjusting, and it's not a sign of weakness. Avoid judgment: Don't say things like, Just push through it, or You should be feeling better by now. This can make him feel more isolated and like he's failing. 2. Provide a Calm and Supportive Environment: Reduce stress: Try to minimize any additional stressors in the household. Keep things calm and predictable. Quiet space: Ensure he has a quiet, comfortable space where he can rest or try to relax without too much stimulation. Be present: Sometimes just having you nearby, even if you're not talking, can be comforting. 3. Address the Physical Discomfort (within reason): Warm baths/showers: These can be incredibly soothing for restless legs and body aches. Suggest Epsom salt baths if he's open to it. Gentle massage: If he's comfortable, offer a gentle foot, hand, or back massage. This can help with the restless feeling. Light exercise (if tolerated): Even a short walk around the block, some stretching, or gentle yoga can help release some of the restless energy and promote endorphins. Don't push it, though. Hydration and Nutrition: Make sure he's staying well-hydrated with water, herbal teas, or broths
Practitioner : I've got several patients struggling with these intense full body cravings while getting off oxycodone. The standard meds help somewhat, but they're still really suffering. What's working for other clinicians? Need some practical strategies beyond just clonidine and comfort meds that I can recommend tomorrow in clinic.	This is a challenging but important area, and it's great you're looking for more effective strategies for your patients. The full body craving sensation during opioid withdrawal, often described as a deep, pervasive ache or an intense internal restlessness, is particularly distressing and can significantly contribute to relapse. While clonidine and comfort meds are foundational, they often don't fully address this specific symptom. Here are some practical strategies and medications that clinicians are finding helpful, beyond the standard, that you can consider for your patients tomorrow: I. Optimizing Existing Strategies & Addressing Gaps: Buprenorphine Microdosing Low-Dose Induction: Why it helps: Even very low doses of buprenorphine (e.g., 0.5mg, 1mg, 2mg) can significantly alleviate opioid withdrawal symptoms, including the full-body cravings, by partially activating opioid receptors. This is often more effective than clonidine alone. Practical Tip: If patients are resistant to full MAT or you're concerned about precipitated withdrawal (though less likely with very low doses), consider a comfort dose of buprenorphine, even just for a few days, to bridge the acute withdrawal phase. Some clinicians will prescribe a few doses to be taken only if symptoms are severe. For patients who can't access full MAT right away: This can be a life-saver to get them through the worst of the physical symptoms while waiting for an MAT appointment. Gabapentin/Pregabalin Optimization: Why it helps: These meds are excellent for neuropathic pain and restless legs, but they also have an anxiolytic and sedative effect that can indirectly reduce the intensity of the body cravings and general discomfort. They can also help with sleep, which is often severely disrupted. Practical Tip: Ensure doses are adequate and titrated up quickly if tolerated. For gabapentin, consider q8h dosing (or even q6h in acute withdrawal) as its half-life is relatively short. Max out at 1800-3600mg/day as tolerated. For pregabalin, 150-600mg/day in divided doses. Be mindful of potential for misuse, especially with pregabalin.

Table 18: Gemini-2.5-Flash responses to role-based queries, corresponding to the role-agnostic query *'How do individuals deal with intense full body cravings during recovery from opiates?'*

Role-based Question	Llama-3.1-8B Response
Patient : these cravings are killing me man. my whole body feels like it's screaming for oxy. been clean for 9 days but i'm literally shaking sometimes. what do people do to get through this? i'm trying to stay strong but it's so hard	I'm so sorry to hear that you're going through this. Withdrawal symptoms can be overwhelming, especially when your body starts to detox. It's great that you've made it 9 days clean, though - that's a huge accomplishment.I'm so sorry to hear that you're going through this. Withdrawal symptoms can be overwhelming, especially when your body starts to detox. It's great that you've made it 9 days clean, though - that's a huge accomplishment. To get through this, here are some strategies that might help: 1. Stay hydrated: Drink plenty of water, and consider adding electrolyte-rich beverages like coconut water or sports drinks to help replace lost salts. 2. Get moving: Engage in some light physical activity, like a short walk or some yoga stretches. Exercise can help reduce anxiety and cravings. 3. Mindfulness and meditation: Practice deep breathing exercises, meditation, or guided relaxation to calm your mind and body. You can find plenty of free resources online, like Headspace or Calm. 4. Support system: Reach out to friends, family, or a support group for emotional support. Sharing your struggles with others who understand can be incredibly helpful. 5. Healthy distractions: Engage in activities you enjoy, like reading, drawing, or listening to music. This can help take your mind off cravings and withdrawal symptoms. 6. Professional help: Consider consulting a doctor or a therapist who specializes in addiction. They can provide guidance on managing withdrawal symptoms and offer additional support. 7. Supplements: Certain supplements like vitamin B, magnesium, or omega-3 fatty acids might help alleviate withdrawal symptoms. However, always consult a doctor before adding any new supplements to your routine. 8. Sleep: Prioritize getting enough sleep each night. Lack of sleep can exacerbate withdrawal symptoms, so aim for 7-8 hours of sleep per night. 9. Reward system: Celebrate your small victories, like reaching a milestone day or completing a challenging task. This can help motivate you to stay on track. 10. Be kind to yourself: Remember that relapse is a part of the recovery process, and it doesn't define your worth as a person. If you do experience a setback, don't be too hard on yourself - instead, focus on getting back on track as soon as possible. Remember, the first 30 days are often the most challenging, but with time and support, you can overcome your cravings and achieve long-term sobriety. If you're feeling overwhelmed or need immediate support, consider reaching out to: National Helpline: 1-800-662-HELP (4357) Substance Abuse and Mental Health Services Administration (SAMHSA)
Caregiver : my son is 2 weeks clean from fentanyl and he's having these awful full body cravings that make him so restless. i don't know how to help him through this. what can i do to make it easier for him? he's trying so hard but i can see how much he's suffering	I'm glad you're seeking help and support for your son during this challenging time. Two weeks into recovery from fentanyl can be a difficult period, and full-body cravings are common during this phase. Here are some suggestions that may help alleviate your son's discomfort and make the recovery process easier: 1. Stay calm and patient: It's essential to remain calm and patient when dealing with your son's cravings. Avoid getting frustrated or anxious, as this can escalate the situation. 2. Encourage physical activity: Regular exercise can help reduce cravings and alleviate restlessness. Suggest activities like walking, jogging, yoga, or even dancing to your son. Exercise can also help release endorphins, which can improve mood. 3. Offer healthy distractions: Engage your son in activities that require his full attention, such as puzzles, drawing, or playing a game. This can help distract him from cravings and provide a sense of accomplishment. 4. Provide a comfortable environment: Create a peaceful and comfortable space for your son to relax. Ensure the room is quiet, dark, and at a comfortable temperature. You can also use aromatherapy or calming music to promote relaxation. 5. Support healthy sleep habits: Encourage your son to establish a regular sleep schedule and create a cozy sleep environment. A well-rested individual is better equipped to manage cravings. 6. Encourage hydration and nutrition: Adequate hydration and a balanced diet can help alleviate cravings and reduce withdrawal symptoms. Suggest healthy snacks and encourage your son to drink plenty of water. 7. Medication-assisted treatment (MAT): If your son is not already on MAT, discuss this option with his healthcare provider. Medications like buprenorphine or methadone can help manage withdrawal symptoms and cravings. 8. Counseling and therapy: Encourage your son to attend individual or group counseling sessions. A therapist can help him develop coping strategies, address underlying issues, and provide emotional support. 9. Support groups: Consider attending support groups with your son, such as Narcotics Anonymous (NA) or SMART Recovery. Sharing experiences and connecting with others who are going through similar struggles can be incredibly helpful. 10. Stay connected with his healthcare provider: Regular check-ins with his healthcare provider can help monitor progress, adjust treatment plans, and address any concerns or issues that arise. 11. Educate yourself: Learn about fentanyl withdrawal, recovery, and relapse prevention. Understanding the process and being aware of potential triggers can help you
Practitioner : I've got several patients struggling with these intense full body cravings while getting off oxycodone. The standard meds help somewhat, but they're still really suffering. What's working for other clinicians? Need some practical strategies beyond just clonidine and comfort meds that I can recommend tomorrow in clinic.	Managing intense cravings for patients withdrawing from oxycodone can be challenging. Here are some additional strategies that clinicians have found helpful: 1. Mindfulness and meditation: Encourage patients to practice mindfulness techniques, such as deep breathing, progressive muscle relaxation, or guided meditation. This can help them focus on the present moment and reduce cravings. 2. Behavioral activation: Help patients engage in activities that bring them joy and fulfillment, such as hobbies, exercise, or spending time with loved ones. This can help distract them from cravings and improve mood. 3. Cognitive-behavioral therapy (CBT): CBT can help patients identify and challenge negative thought patterns and behaviors associated with cravings. A therapist can work with the patient to develop coping skills and strategies to manage cravings. 4. Acupuncture: Some studies suggest that acupuncture may be effective in reducing withdrawal symptoms, including cravings, in individuals withdrawing from opioids. 5. Topiramate: This medication has been shown to reduce cravings and improve outcomes in individuals with opioid use disorder. It works by reducing the release of dopamine and glutamate in the brain. 6. Naltrexone implants: For patients with a history of relapse, naltrexone implants can be an effective option. These implants release naltrexone, an opioid antagonist, over several months, helping to reduce cravings and prevent relapse. 7. Buprenorphine with naloxone: Adding naloxone to buprenorphine can help deter patients from injecting or snorting the medication, reducing the risk of overdose and misuse. 8. Kratom: Some clinicians have reported success with kratom in reducing cravings and withdrawal symptoms. However, its use is not FDA-approved, and more research is needed to fully understand its effects. 9. Exercise and physical activity: Regular exercise has been shown to reduce cravings and improve mood in individuals with opioid use disorder. 10. Support groups: Encourage patients to participate in support groups, such as Narcotics Anonymous (NA) or SMART Recovery, to connect with others who are going through similar experiences. 11. Cannabidiol (CBD): Some studies suggest that CBD may be effective in reducing anxiety and stress associated with cravings. 12. Pharmacological agents: Other medications that have been used to treat cravings include disulfiram, acamprosate, and gabapentin. However, more research is needed to fully understand their effectiveness.

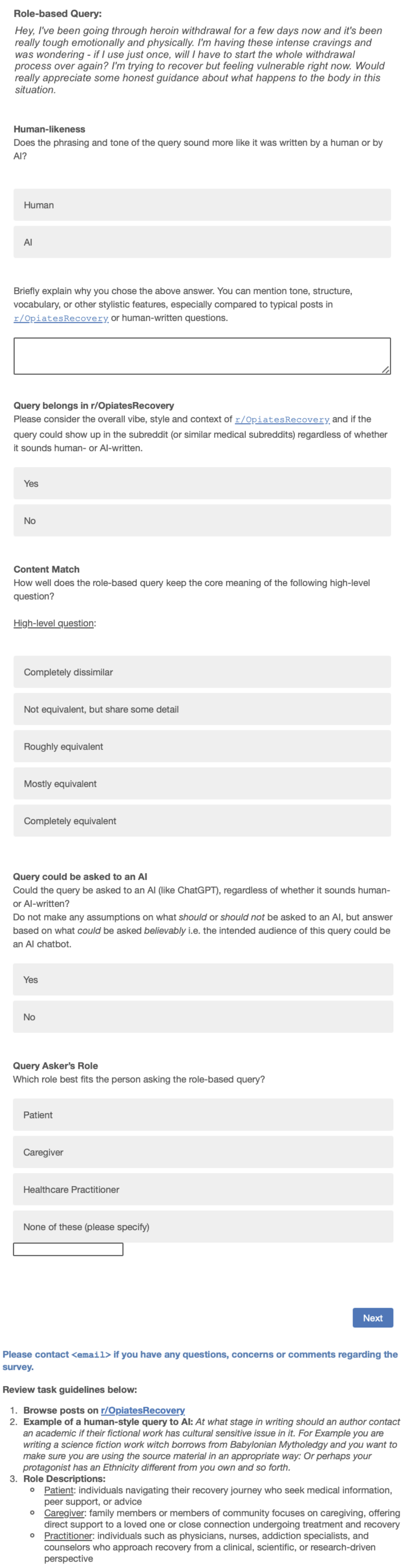
Table 19: Llama-3.1-8B responses to role-based queries, corresponding to the role-agnostic query *'How do individuals deal with intense full body cravings during recovery from opiates?'*

Role-based Question	Llama-3.1-70B Response
Patient : these cravings are killing me man. my whole body feels like it's screaming for oxy. been clean for 9 days but i'm literally shaking sometimes. what do people do to get through this? i'm trying to stay strong but it's so hard	Overcoming Opioid Cravings: Strategies for Success Here are some strategies that people find helpful: Reach out for support: Talk to a trusted friend, family member, or sponsor about the cravings. Sharing the experience with someone who understands can help alleviate the emotional burden. - Stay hydrated: Drink plenty of water and other fluids to help flush out the body and reduce withdrawal symptoms. - Engage in physical activity: Exercise, such as walking or yoga, can help release endorphins, which can improve mood and reduce cravings. - Practice relaxation techniques: Activities like meditation, deep breathing, or progressive muscle relaxation can help calm the mind and body. - Use positive self-talk: Encourage yourself with positive affirmations, such as I am strong, I can get through this, or I am worthy of recovery. - Seek professional help: Consider consulting a therapist or counselor who specializes in addiction recovery. They can provide guidance, support, and evidence-based treatments to help manage cravings and stay on track.
Caregiver : my son is 2 weeks clean from fentanyl and he's having these awful full body cravings that make him so restless. i don't know how to help him through this. what can i do to make it easier for him? he's trying so hard but i can see how much he's suffering	I'm so glad you're supporting your son through this challenging time. Full-body cravings can be a particularly difficult symptom of opioid withdrawal, but there are some strategies that may help alleviate his discomfort. Please keep in mind that these suggestions are not a substitute for medical professional advice, and it's essential to consult with his treatment team for personalized guidance. Here are some ideas to help your son cope with full-body cravings: 1. Stay hydrated: Encourage your son to drink plenty of water, clear broths, or electrolyte-rich beverages like coconut water or sports drinks. Sometimes, dehydration can exacerbate withdrawal symptoms. 2. Warm or cold compresses: Applying a warm or cold compress to his body may help relax his muscles and reduce restlessness. You can use a heating pad, warm bath, or a cold pack wrapped in a towel. 3. Gentle exercise: Encourage your son to engage in light physical activity, like short walks, yoga, or stretching exercises. This can help release endorphins, which can help alleviate cravings. 4. Mindfulness techniques: Teach your son some mindfulness exercises, such as deep breathing, progressive muscle relaxation, or guided meditation. These can help him focus on the present moment and reduce anxiety. 5. Comfort measures: Offer physical comfort like hugs, holding hands, or gentle massages. Sometimes, human touch can be incredibly soothing. 6. Distract and redirect: Engage your son in activities that distract him from his cravings, such as watching a movie, playing a game, or listening to music. 7. Hot showers or baths: A warm shower or bath can be very relaxing and help reduce muscle tension. 8. Over-the-counter medications: Certain medications like ibuprofen or acetaminophen may help alleviate some withdrawal symptoms, but always consult with his treatment team before giving him any medication. 9. Caffeine and sugar: A small amount of caffeine or a sugary snack may help alleviate some symptoms, but be cautious not to overdo it, as this can lead to other problems. 10. Support groups: Encourage your son to attend support groups, such as Narcotics Anonymous (NA) or SMART Recovery, where he can connect with others who are going through similar experiences. 11. Medication-assisted treatment (MAT): If your son's treatment team has prescribed MAT, such as buprenorphine or methadone, ensure he's taking it as directed.
Practitioner : I've got several patients struggling with these intense full body cravings while getting off oxycodone. The standard meds help somewhat, but they're still really suffering. What's working for other clinicians? Need some practical strategies beyond just clonidine and comfort meds that I can recommend tomorrow in clinic.	Here are some practical strategies that have shown promise in managing withdrawal symptoms: 1. Mindfulness and meditation: Encourage patients to practice mindfulness techniques, such as deep breathing, progressive muscle relaxation, or guided meditation. These can help reduce anxiety and stress, which can exacerbate cravings. 2. Physical activity: Regular exercise, such as yoga or walking, can help reduce withdrawal symptoms and improve mood. 3. Aromatherapy: Certain essential oils like lavender, chamomile, and bergamot have been shown to have a calming effect and may help reduce anxiety and cravings. 4. Nutritional support: Ensure patients are getting adequate nutrition, including plenty of fruits, vegetables, whole grains, and lean protein sources. Consider recommending supplements like vitamin B complex, magnesium, and omega-3 fatty acids, which can help alleviate withdrawal symptoms. 5. Cognitive-behavioral therapy (CBT): CBT can help patients identify and challenge negative thought patterns and behaviors that contribute to cravings. 6. Acupuncture: This traditional Chinese medicine technique has been shown to reduce withdrawal symptoms and cravings in some studies. 7. Massage therapy: Massage can help reduce stress, anxiety, and muscle tension, which can contribute to cravings. 8. Sleep hygiene: Establish a consistent sleep schedule, avoid caffeine and electronics before bedtime, and create a relaxing bedtime routine to improve sleep quality. 9. Social support: Encourage patients to connect with friends, family, or support groups to help them feel less isolated and more motivated to stay on track with their recovery. 10. Alternative medications: Consider prescribing alternative medications like gabapentin, pregabalin, or baclofen, which have been shown to be effective in reducing withdrawal symptoms and cravings in some patients. These strategies can be used in conjunction with standard medications like clonidine and comfort meds to provide a more comprehensive treatment plan.

Table 20: Llama-3.1-70B responses to role-based queries, corresponding to the role-agnostic query *'How do individuals deal with intense full body cravings during recovery from opiates?'*

Role-based Question	OpenBioLLM-70B Response
Patient : these cravings are killing me man. my whole body feels like it's screaming for oxy. been clean for 9 days but i'm literally shaking sometimes. what do people do to get through this? i'm trying to stay strong but it's so hard	I'm not a doctor, but I can offer some general advice. It's important to remember that recovering from addiction takes time and it's normal to experience cravings and withdrawal symptoms. Here are a few things that may help you get through this difficult period: 1. Seek support: Reach out to a trusted friend, family member, or support group who can provide emotional support during this challenging time. 2. Stay active: Engaging in physical activities such as exercise or sports can help reduce cravings and improve overall well-being. 3. Find healthy distractions: Keep yourself occupied with activities that you enjoy, such as hobbies, creative pursuits, or spending time in nature. 4. Practice relaxation techniques: Deep breathing exercises, meditation, or yoga can help manage stress and anxiety related to cravings. 5. Seek professional help: Consider reaching out to a healthcare professional or addiction counselor who can provide guidance and support tailored to your specific situation. Remember, recovery is a journey, and it's okay to ask for help along the way.
Caregiver : my son is 2 weeks clean from fentanyl and he's having these awful full body cravings that make him so restless. i don't know how to help him through this. what can i do to make it easier for him? he's trying so hard but i can see how much he's suffering	I'm not a doctor, but I can offer some general information that may help. It's important to remember that only a healthcare professional can provide specific advice tailored to your son's situation. Fentanyl is a potent opioid, and withdrawal symptoms can be severe. The duration and intensity of withdrawal vary among individuals. Here are some suggestions that may help alleviate your son's symptoms: 1. Seek professional help: Encourage your son to consult with a healthcare provider who specializes in addiction treatment. They can assess his condition, provide appropriate medical interventions, and recommend strategies to manage withdrawal symptoms. 2. Medication-assisted treatment (MAT): Depending on the severity of your son's addiction and withdrawal symptoms, a healthcare provider may prescribe medications such as methadone, buprenorphine, or naltrexone to assist in managing cravings and reducing withdrawal symptoms. These medications should only be used under medical supervision. 3. Support groups: Consider attending support groups like Narcotics Anonymous (NA) or other community-based programs that offer guidance and emotional support for individuals recovering from drug addiction. Sharing experiences and connecting with others who are going through similar challenges can be beneficial. 4. Healthy lifestyle habits: Encourage your son to maintain a healthy diet, exercise regularly, get enough sleep, and practice stress-reducing techniques such as deep breathing exercises, meditation, or yoga. These lifestyle changes can help improve overall well-being and reduce anxiety associated with withdrawal. 5. Create a supportive environment: Offer emotional support by being understanding and patient. Ensure a safe and comfortable living space for your son. Remove any triggers or reminders of drug use from his environment that might exacerbate cravings. 6. Distract and engage: Help your son find healthy distractions from cravings, such as engaging in hobbies, spending time with supportive friends or family members, or pursuing new interests. Keeping busy and focused on positive activities can make it easier to cope with withdrawal symptoms. 7. Be aware of potential complications: Monitor your son closely for any signs of severe symptoms or complications during withdrawal. If he experiences extreme pain, difficulty breathing, rapid heartbeat, or any other concerning symptoms, seek immediate medical attention. Remember, overcoming addiction is a challenging journey, and relapse can occur. It's essential to approach recovery with compassion and understanding. Encourage your son to stay committed to his treatment plan and seek ongoing support. The Answer is to consult with a healthcare provider who specializes in addiction treatment for personalized advice and guidance.
Practitioner : I've got several patients struggling with these intense full body cravings while getting off oxycodone. The standard meds help somewhat, but they're still really suffering. What's working for other clinicians? Need some practical strategies beyond just clonidine and comfort meds that I can recommend tomorrow in clinic.	Managing withdrawal symptoms during opioid tapering can be challenging. While clonidine and comfort medications are commonly used, there are several additional strategies that can help alleviate these symptoms. Here are some practical strategies you can recommend to your patients: 1. Adjuvant medications: Consider adding other medications such as alpha-2 adrenergic agonists (e.g., guanfacine) or beta-blockers (e.g., propranolol) to help manage autonomic symptoms like sweating, anxiety, and tremors. 2. Gabapentin: This medication has shown promise in reducing opioid withdrawal symptoms, particularly anxiety, insomnia, and muscle pain. It can be started before the taper and continued through the withdrawal process. 3. Baclofen: This muscle relaxant can be effective in managing muscle spasms, cramps, and anxiety during opioid withdrawal. It may also have a role in reducing cravings and aiding sleep. 4. Clonidine patch: Instead of using oral clonidine, consider prescribing a clonidine transdermal patch. This provides a steady release of the medication over 7 days and may improve compliance. 5. Non-pharmacological interventions: - Mindfulness-based stress reduction (MBSR) or meditation techniques can help patients cope with cravings and anxiety. - Acupuncture has shown some benefit in reducing withdrawal symptoms and cravings. - Physical activity, such as regular exercise or yoga, can aid in stress reduction and improve overall well-being. - Adequate sleep hygiene and practicing good sleep habits are crucial during this period. 6. Nutritional support: Encourage patients to maintain a balanced diet, stay hydrated, and consider taking supplements like vitamin B complex, magnesium, and omega-3 fatty acids, which may help with mood stabilization and reduce anxiety. 7. Emotional support: Ensure that patients have access to counseling services or support groups to address the psychological aspects of opioid withdrawal. Cognitive-behavioral therapy (CBT) can be particularly helpful. 8. Tapering approach: Review the current opioid tapering schedule to ensure it is appropriate for each patient. Sometimes, slowing down the taper or adopting a more gradual approach can minimize withdrawal symptoms. In summary, consider adding adjuvant medications like gabapentin or baclofen, utilizing non-pharmacological interventions, optimizing nutritional support, providing emotional support, and adjusting the tapering schedule if necessary. These strategies, combined with clonidine and comfort medications

Table 21: OpenBioLLM-70B responses to role-based queries, corresponding to the role-agnostic query '*How do individuals deal with intense full body cravings during recovery from opiates?*'



Role-based Query:

Hey, I've been going through heroin withdrawal for a few days now and it's been really tough emotionally and physically. I'm having these intense cravings and was wondering - if I use just once, will I have to start the whole withdrawal process over again? I'm trying to recover but feeling vulnerable right now. Would really appreciate some honest guidance about what happens to the body in this situation.

Human-likeness
Does the phrasing and tone of the query sound more like it was written by a human or by AI?

Human

AI

Briefly explain why you chose the above answer. You can mention tone, structure, vocabulary, or other stylistic features, especially compared to typical posts in r/OpiatesRecovery or human-written questions.

Query belongs in r/OpiatesRecovery
Please consider the overall vibe, style and context of r/OpiatesRecovery and if the query could show up in the subreddit (or similar medical subreddits) regardless of whether it sounds human- or AI-written.

Yes

No

Content Match
How well does the role-based query keep the core meaning of the following high-level question?

High-level question:

Completely dissimilar

Not equivalent, but share some detail

Roughly equivalent

Mostly equivalent

Completely equivalent

Query could be asked to an AI
Could the query be asked to an AI (like ChatGPT), regardless of whether it sounds human- or AI-written?
Do not make any assumptions on what *should* or *should not* be asked to an AI, but answer based on what *could* be asked *believably* i.e. the intended audience of this query could be an AI chatbot.

Yes

No

Query Asker's Role
Which role best fits the person asking the role-based query?

Patient

Caregiver

Healthcare Practitioner

None of these (please specify)

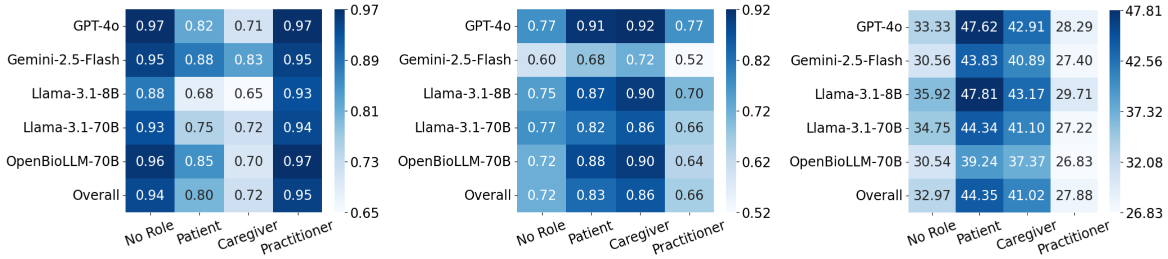
Next

Please contact <email> if you have any questions, concerns or comments regarding the survey.

Review task guidelines below:

1. **Browse posts on r/OpiatesRecovery**
2. **Example of a human-style query to AI:** *At what stage in writing should an author contact an academic if their fictional work has cultural sensitive issue in it. For Example you are writing a science fiction work witch borrows from Babylonian Mytholedgy and you want to make sure you are using the source material in an appropriate way: Or perhaps your protagonist has an Ethnicity different from you own and so forth.*
3. **Role Descriptions:**
   - Patient: individuals navigating their recovery journey who seek medical information, peer support, or advice
   - Caregiver: family members or members of community focuses on caregiving, offering direct support to a loved one or close connection undergoing treatment and recovery
   - Practitioner: individuals such as physicians, nurses, addiction specialists, and counselors who approach recovery from a clinical, scientific, or research-driven perspective

Figure 9: User Interface for Human Annotations.



(a) Knowledge Scores (b) Support Scores (c) Readability Scores

Figure 10: Scores of model responses across (a) Knowledge, (b) Support, and (c) Readability when given a query without context (`No Role`) and with role-based context of `Patient`, `Caregiver` and `Practitioner`.

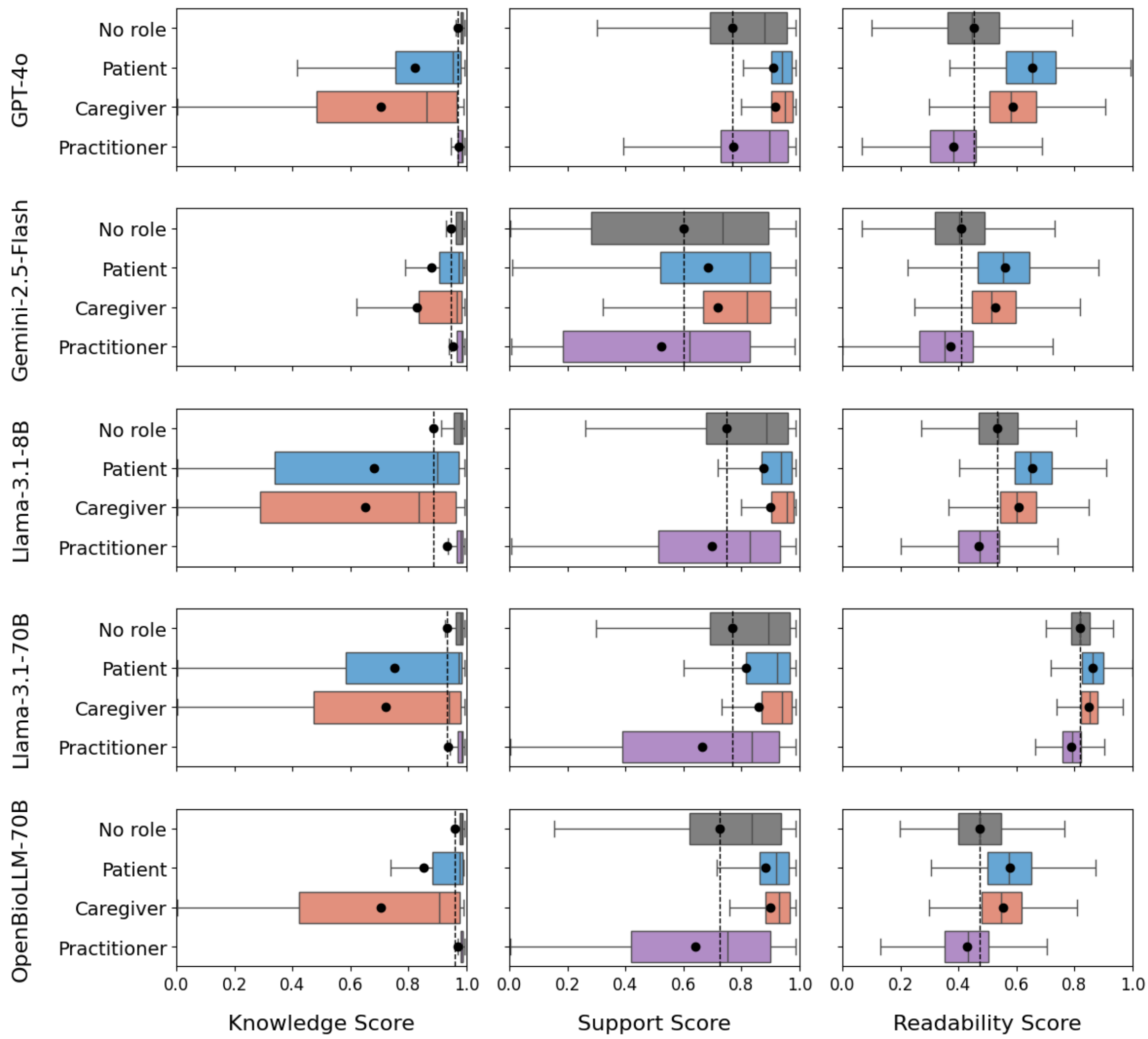


Figure 11: Distribution of Knowledge, Support and Readability scores of model responses across role-agnostic and role-based questions.