"What's Up, Doc?": Analyzing How Users Seek Health Information in Large-Scale Conversational AI Datasets

Akshay Paruchuri[†] **Maryam Aziz**[◇] **Rohit Vartak**[◇] **Ayman Ali**[◇]
Best Uchehara[◇] **Xin Liu**[○•] **Ishan Chatterjee**[○•] **Monica Agrawal**[◇]
[†] UNC Chapel Hill [◇] Duke University [○] University of Washington [•] Google
akshay@cs.unc.edu monica.agrawal@duke.edu

Abstract

People are increasingly seeking healthcare information from large language models (LLMs) via interactive chatbots, yet the nature and inherent risks of these conversations remain largely unexplored. In this paper, we filter large-scale conversational AI datasets to achieve HealthChat-11K, a curated dataset of 11K real-world conversations composed of 25K user messages. We use HealthChat-11K and a clinician-driven taxonomy for how users interact with LLMs when seeking healthcare information in order to systematically study user interactions across 21 distinct health specialties. Our analysis reveals insights into the nature of how and why users seek health information, such as common interactions, instances of incomplete context, affective behaviors, and interactions (e.g., leading questions) that can induce sycophancy, underscoring the need for improvements in the healthcare support capabilities of LLMs deployed as conversational AI. Code and artifacts to retrieve our analyses and combine them into a curated dataset can be found here: https://github.com/yahskapar/HealthChat

1 Introduction

Large language models (LLMs) have demonstrated significant medical knowledge which has translated to proficiency at a range of clinical tasks including differential diagnosis, interpretation of health records, and medical summarization (Agrawal et al., 2022; Singhal et al., 2023; Thirunavukarasu et al., 2023; McDuff et al., 2025). These impressive capabilities, coupled with the costs and inaccessibility of traditional medical care, have led a growing number of people to turn to LLM-based chatbots to seek healthcare information. One 2024 survey found that 31% of US adults were turning to generative AI for health requests, seeking help with self-diagnosis, treatment management, and support needs (Choy et al., 2024).

Despite this surge in public use, the vast majority of evaluations of LLMs in medicine rely on benchmarks that focus on clinician- or researcher-oriented tasks (Raji et al., 2025; Bedi et al., 2024). These benchmarks often assume a structured, professional context that differs substantially from how patients engage with chatbots in the real world. While public datasets focused on consumer health queries do exist, they take the form of single-turn health queries or synthetic interactions (Arora et al., 2025; Kilicoglu et al., 2018; Singhal et al., 2023).

Unfortunately, these existing benchmarks provide a suboptimal proxy for the open-ended, often ambiguous questions posed by lay users over the course of multi-turn conversations. A recent large-scale user study found that the clinical knowledge measured by current synthetic benchmarks is insufficient to account for the failure modes surfaced via real human interactions (Bean et al., 2025). Additional recent work has shown that LLMs are suboptimal at soliciting further details when only partial information is provided for differential diagnosis (Zhao et al., 2024; Johri et al., 2025).

Furthermore, LLMs are known to exhibit problematic tendencies such as sycophancy, overconfidence, and hedging, which can seriously impact the quality and reliability of healthcare information given to users (Sharma et al., 2023; Ranaldi and Pucci, 2023; Yang et al., 2024; Yona et al., 2024). Given these known limitations of LLMs, there is an urgent need to characterize real-world communication patterns of people querying chatbots for healthcare information to understand these behaviors (and their corresponding risks) under realistic conditions.

In this work, we investigate and systematically analyze common interactions, instances of providing incomplete context, affective behaviors, and interactions (e.g., leading questions) that can induce sycophancy by real-world users engaging with LLM-based chatbots for healthcare information.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2312–2336
November 4-9, 2025

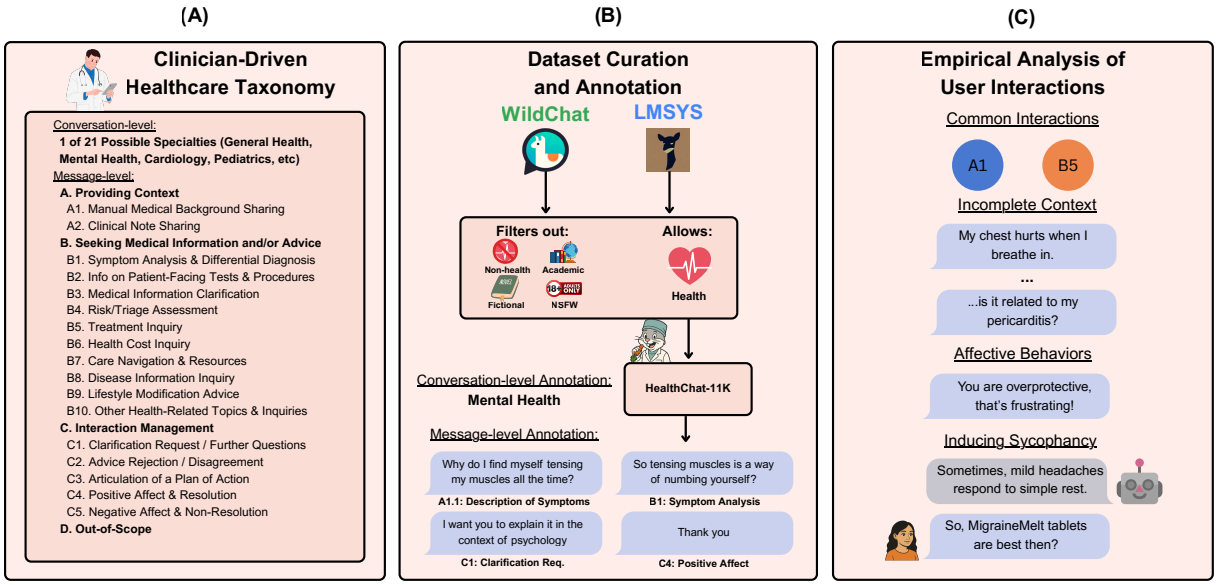

Figure 1: **Overview.** In collaboration with clinicians, we **(A)** develop a clinician-driven healthcare interaction taxonomy in order to **(B)** classify and annotate a curated dataset of conversations filtered from large-scale conversational datasets such as LMSYS-Chat-1M (Zheng et al., 2023) and WildChat-1M (Zhao et al., 2024). An LLM is used to apply conversation-level specialty annotations and message-level taxonomic code annotations to the curated dataset. Subsequently, we leverage our annotations to **(C)** analyze users' common interactions, instances of providing incomplete context, affective behaviors, and interactions (e.g., leading questions) that can induce sycophancy.

Our contributions are as follows:

1. Section 3: We filter conversations from large-scale conversational datasets such as LMSYS-Chat-1M (Zheng et al., 2023) and WildChat-1M (Zhao et al., 2024) to identify 11K real-world conversations composed of 25K user messages.

2. Section 4: We develop and apply a clinician-driven taxonomy for how users interact with LLMs when seeking healthcare information and classify conversations into one of 21 distinct health specialties.

3. Section 5 and Section 6: Finally, through our clinician-driven healthcare interaction taxonomy, we analyze the dataset, investigating users' common interaction patterns, as well as instances of patients providing incomplete context, displaying affective behaviors, and potentially inducing sycophancy.

To foster future research, we have released code and artifacts to retrieve our analyses and combine them into a curated dataset here: https://github.com/yahskapar/HealthChat

2 Related Work

Online Health Information Searching There is a significant imbalance between clinician availability and patient needs (for Health Workforce Analysis, 2024). In response, patients have long turned to the Internet, e.g. search engines and symptom checkers, to independently seek health information (Wang et al., 2021). Although these tools can be useful, they can also misinform patients (even unintentionally), with some patients placing greater trust in online information than their providers (Davis, 2018). More recently, general-purpose generative AI tools, such as chatbots, have emerged as a new avenue for seeking health information (Choy et al., 2024). However, beyond survey-level insights, we know little about how patients engage with these AI systems in multi-turn interactions or how such behavior compares to earlier modes of online information search (Chang et al., 2022). The potential for LLMs to dramatically improve information accessibility coupled with their potential harms highlight an urgent need to better understand the nature of these interactions, as we do through our clinician-driven healthcare interaction taxonomy.

Datasets of Patient Information Needs There are comparatively fewer datasets focused on patients' health information needs, as most evaluations of LLMs in healthcare are targeted from a clinician-centric perspective (Bedi et al., 2024). Existing real-world datasets like CHQA, ChatDoctor,

and HealthSearchQA are single-turn, sourced from online patient forums or web searches (Kilicoglu et al., 2018; Singhal et al., 2023; Li et al., 2023). Existing multi-turn datasets are synthetically generated, limiting their utility for studying real-world patients' interaction patterns (Arora et al., 2025).

Conversational Datasets The growing popularity of LLMs (Achiam et al., 2023; Team et al., 2023) as conversational knowledge interfaces has prompted further research and development toward analyzing and improving conversational AI capabilities. Large-scale, real-world conversation datasets such as LMSYS-Chat-1M (Zheng et al., 2023), WildChat-1M (Zhao et al., 2024), and ShareGPT (Wang et al., 2023) jointly consist of millions of diverse human-AI conversations. They have been used both for open-domain benchmarking and instruction-tuning of models (Chiang et al., 2023). In contrast to raw conversations, alignment research, like the PRISM dataset (Kirk et al., 2024), examines who provides feedback and why, using demographic data to enable personalized AI.

User Interaction and Behavior Dynamics with Conversational AI. The increasing reliance on Large Language Models (LLMs) for healthcare advice necessitates a deeper understanding of user interaction dynamics. Taxonomies have been developed and applied in order to analyze conversational agents that utilize natural language when interacting with users (Feine et al., 2019). This has extended from more classical agents that utilize natural language processing to knowledgeable AI agents that are capable of human-AI collaboration (Dellermann et al., 2021). Similarly, in clinical settings, health researchers have developed taxonomies to capture patterns in patient complaints to reveal latent safety and quality problems across clinical systems (Reader et al., 2014). In non-clinical settings, researchers have built benchmarks to understand multi-turn dialogue quality (Bai et al., 2024), identified critical failures in conversational grounding (Shaikh et al., 2025), and determined that users often prioritize functional substance over conversational style when conversing with conversational coaching agents (Srinivas et al., 2025). Researchers have also introduced a taxonomy that maps ChatGPT's functions and risks within clinical, educational, and administrative workflows, underscoring a particular need for structured evaluation of LLM behaviour in medical contexts (Li et al., 2024a). In contrast to (Li et al., 2024a), our work with HealthChat-11K shifts the focus from a high-level review of research applications to a direct, empirical analysis of user interactions. Our clinician-driven taxonomy is applied at the message-level across 11K real-world conversations to map user interactions and subsequently provides a user-centric view of how health discussions with LLMs can unfold in the real-world. Lastly, past research has uncovered potentially undesirable language model behaviors, such as sycophancy (Sharma et al., 2023), which we study through the lens of user interactions while seeking treatment recommendations.

3 Dataset Creation

We first curate a dataset of health information seeking conversations and then apply conversation-level and message-level taxonomic annotations in a scalable manner. Our dataset curation process consists of several distinct steps - 1) we use labels from both WildChat-1M and LMSYS-Chat-1M to filter out any non-English, toxic conversations. Then, 2) we create and apply LLM-based filtering prompts to filter out non-health conversations using Gemini 1.5 Pro. 20 conversations are evaluated at a time and subsequent refinement of the prompt occurs after manual inspection of filtered conversations. The initial, LLM-based filtering prompt targeting the removal of non-health, academic, fictional, and not-safe-for-work (NSFW) conversations brings us to approx. 37K conversations, after which 3) another, LLM-based filtering prompt with few-shot examples is applied in order to more explicitly reject conversations that were observed to be undesirable through empirical observation.

HealthChat Characteristic	**Value**
Total Conversations	11K
Total User Messages	25K
Average User Message Length (tokens)	22
Conversation Turn Distribution:	
Conversations with 1 Turn	*56%*
Conversations with 2 Turns	*18%*
Conversations with 3 Turns	*10%*
Conversations with 4+ Turns	*16%*

Table 1: **HealthChat Dataset Characteristics.**

Following the aforementioned LLM-based filtering prompt with few-shot examples, we end up with approx. 16K conversations. We additionally

4) manually mark and filter out approx. 400 conversations and utilize a pre-trained Sentence Transformer model (Reimers and Gurevych, 2019), all-MiniLM-L6-v2, to filter out approx. 100 additional conversations that had too high of an embedding similarity (≥ 0.9) to undesirable conversation examples. Then, 5) we perform de-duplication to remove an additional 2000 conversations. After proceeding with taxonomic annotation that is further detailed in Section 4.3, we also take one final step to curate our dataset: 6) utilizing D (Out-of-Scope) codes to filter our dataset. After dropping any conversations where more than half of the user messages include D codes, we end up with a final dataset of 11K real-world conversations and 25K user messages. Further details (e.g., prompts, hyperparameters) of our dataset curation process can be found in Appendix A. Our final dataset statistics can be found in Table 1.

Human-LLM Concordance. By comparing the inclusion or exclusion decisions of an expert annotator (author of this paper with expertise in public health) in contrast to the LLM (Gemini 1.5 Pro) on a test set of 300 conversations, we observe a precision of 0.85 when it came to inclusion concordance. A rubric specifying the human annotation criteria can be found in Appendix A.3.

4 A Taxonomy for Conversations Seeking Health Information

We develop and apply a taxonomy to (i) allow for more robust identification and categorization of conversations where users are genuinely seeking healthcare advice amidst broader, open-domain interactions, and (ii) enable fine-grained analysis of users' interactions with LLMs for health-related concerns. Our healthcare interaction taxonomy was built in collaboration with two clinicians; an abridged version can be found in Figure 1 and the full version can be found in Appendix B.1.

4.1 Specialty Taxonomy

At the conversation-level, our healthcare interaction taxonomy classifies the overarching health specialty that is present. In collaboration with our clinician collaborators, we curated a list of 21 possible specialties based on the American Board of Medical Specialties list of specialty and subspecialty certificates (American Board of Medical Specialties, 2025). Our list of health specialties includes both more general, broad categories (e.g., general health) and more fine-grained categories (e.g., hematology/oncology, cardiology). A distribution of all 21 specialties can be found in Figure 2.

4.2 Conversation Taxonomy

In addition to conversation-level specialty classification, we utilize four distinct categories at the message-level - A) Providing Context, B) Seeking Medical Information and/or advice, C) Interaction Management, and D) Out-of-Scope. Each of these categories are then subdivided with further granularity. For example, within A) Providing Context, an additional level of granularity is A1) Manual Medical Background Sharing, which further contains several subdivisions such as A1.1) description of relevant acute symptoms, A1.2) sharing of relevant chronic condition(s) and past procedure history, and A1.3) sharing of lab values, or findings from imaging/culture/diagnostic procedures.

As an example, the message *"What is a vasectomy?"* would be annotated with B2 (Information on Patient-Facing Tests and Procedures). A single message can also have multiple taxonomic annotations. For example, the message *"I feel dizzy, I have runny nose, I can't breathe, I'm bored, I'm not hungry, what's wrong?"* includes a description of the relevant acute symptoms (A1.1) and an inquiry around the cause of these symptoms (B1). The codes in C (Interaction Management) cover user modes like clarifications, rejections, plans of actions, and affective behaviors. For example, *"Were you suggesting a mastectomy?"* falls under a clarification request (C1). Lastly, D corresponds to "Out-of-Scope" messages, indicating off-task messages in the conversation unrelated to health seeking information. For our final dataset, we dropped conversations where more than half of the user messages were classified with D codes.

4.3 Taxonomic Annotation

Our taxonomic annotation process leverages an LLM (Gemini 2.5 Pro) in order to annotate taxonomy codes. As per the prompt instructions and few-shot examples, the LLM performs 1) specialty annotation at the conversation-level and 2) taxonomy code annotation at the message level. A single specialty must be assigned at the conversation-level. At the message-level, multiple codes can be assigned to a single message, but every message must have a code and D codes are generally used whenever there is uncertainty or as a last resort. We include further details (e.g., prompts, hyper-

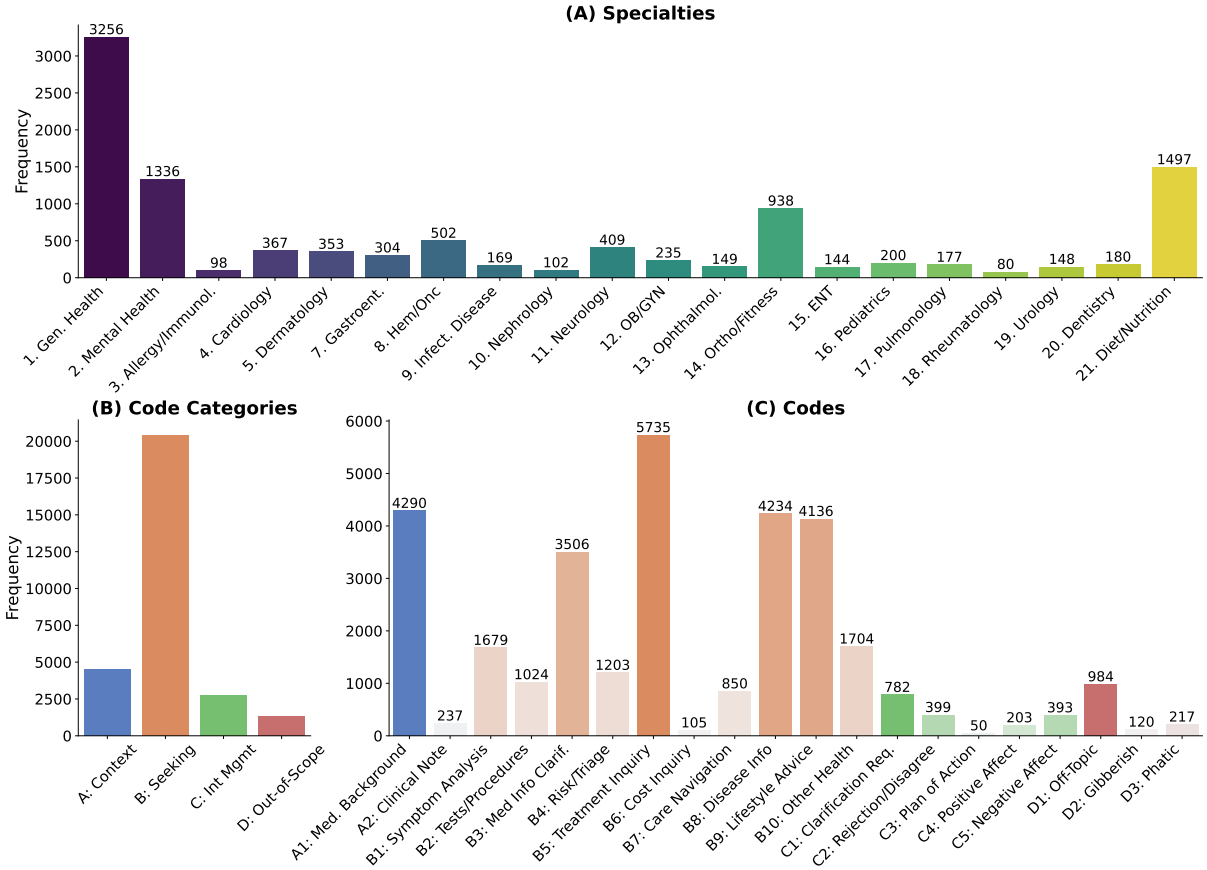

Figure 2: **Specialties and Codes Distributions.** Though certain aspects of our taxonomy (e.g., general health questions and requests for information) appear in higher frequencies as per our taxonomy design and subsequent annotation, reasonable coverage across our proposed taxonomy design is observed. This shows that HealthChat-11K is capable of additional downstream analyses motivated by specific specialties and/or user behaviors (e.g., context disclosure versus seeking information).

parameters) for our taxonomic annotation process in Appendix B.

Human-LLM Concordance. At the second level of conversation taxonomic depth (e.g., A1, B2), the macro-averaged F1-score between Gemini 2.5 Pro and each human annotator (3 total) averaged 0.78, comparable to the inter-annotator F1-score of 0.77 among humans. Krippendorff's alpha was 0.61, indicating moderate agreement and reflecting the challenge of achieving a perfect match between annotation sets for unconstrained and potentially complex user messages seeking health information. The annotation rubric and LLM prompts are detailed in Appendix B.5. An expert annotator evaluated all individual taxonomy codes (for both specialty and conversation labels) via a random sampling of 10 LLM annotations per possible annotation. All codes were verified to have at least 80% accuracy (8/10) as per the sampled annotation examples.

5 Taxonomic Analysis

The creation of our taxonomy in Section 4 in collaboration with clinicians and subsequent annotation (Section 4.3) enables us to analyze common user interactions while seeking health information. Additionally, we analyze the long tail of health information seeking interactions through B10 - a taxonomy code that targets health-related inquiries not covered by other information request (B) codes.

5.1 Common Interactions

Following our taxonomic annotation of 25K user messages, common interactions emerge as shown in Figure 2. In particular, users seek health information where the overarching conversation can be classified into health specialties, most notably general health, mental health, and diet/nutrition. It is also notable that users have almost four times as many interactions seeking information (code category B) than any other kind of interaction. This corresponds to the idea of language models being capable of functioning as factual knowledge

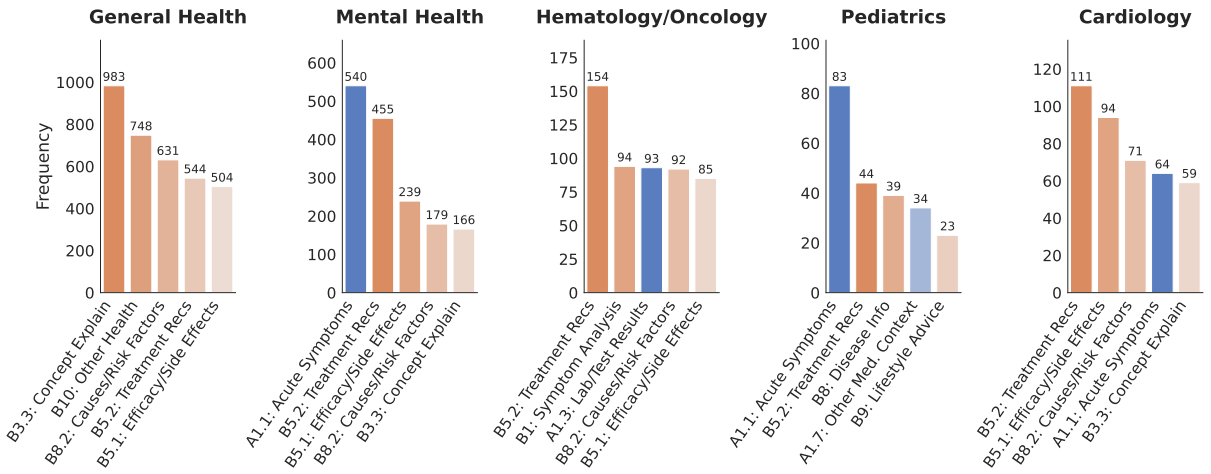

Figure 3: **Distribution of Top 5 Taxonomy Codes in Select Specialties.** User interactions seeking information (e.g., inquiring about disease information) mostly dominate specialty-specific taxonomy code distributions. A notable exception is the *Mental Health* specialty, where users were observed to frequently provide context about their principal concern (i.e., acute symptoms) in order to identify a mental health problem.

bases (Petroni et al., 2019), potentially encouraging users to use them as such even through a chatbot interface. This is further reinforced by quantitative analysis of bi-grams (e.g., $A1 \rightarrow B5$), where it is observed that self-loops (e.g., $B5 \rightarrow B5$) compose almost half of all bi-grams. Almost 40% of self-loops involve requests of information (B codes), and there is a 57% chance that a user message assigned a B code is followed by another user message with a B code. Within the distribution of interactions involving requests for information, treatment inquiry (B5) and lifestyle advice (B9) are most prevalent, and this corresponds to the most common B code self-loops as well. The dominance of these information-seeking patterns could indicate a need for LLMs to evolve from passive knowledge bases into more proactive conversational partners that can carefully solicit more complete information from users in order to ensure appropriate provision of information in sensitive areas such as healthcare support.

The theme of users predominantly seeking information remains true even when we observe select specialties, as shown in Figure 3. Manual medical background sharing interactions (A1) appear in four of five selected specialties and their corresponding distributions of the top five taxonomy codes, with relatively higher, or comparable to information seeking, occurrences of this context providing behavior occurring in the specialties of mental health and pediatrics. In the case of mental health in particular, users were observed to frequently provide context about their acute symptoms in order to identify a mental health problem. This indicates that with mental health in particular, and perhaps to a lesser extent with pediatrics, a more common user interaction involves providing context in order to deal with acute symptoms and identify a health problem in the first place. Identifying mental health problems can be tricky in clinical settings to begin with because diagnosis relies on subjective self-reporting rather than objective biomarkers (APA, 2013), and it makes sense that this can be even more challenging via LLM-based chatbots because text-based interactions lack the empathetic probing and diagnostic insight a human clinician can provide (Lawrence et al., 2024).

5.2 Examining "Other Health Requests"

Our taxonomy was guided by anticipated common health information needs, but it is important to understand the long tail of health information seeking. To do so, we further examined a set of 100 random responses that occurred in the "B10: Other Health Request" category, which comprises approximately 7% of the dataset. Recurring themes included asking about AI in medicine (e.g., how AI-assisted diagnosis works, the ethics of using chatbots for health advice); the practice and legality of certain health practices (e.g., treatments, procedures) in different countries; the necessity and harms of vaccines and masking; general overviews of a field (e.g., nutrition); and for population trends.

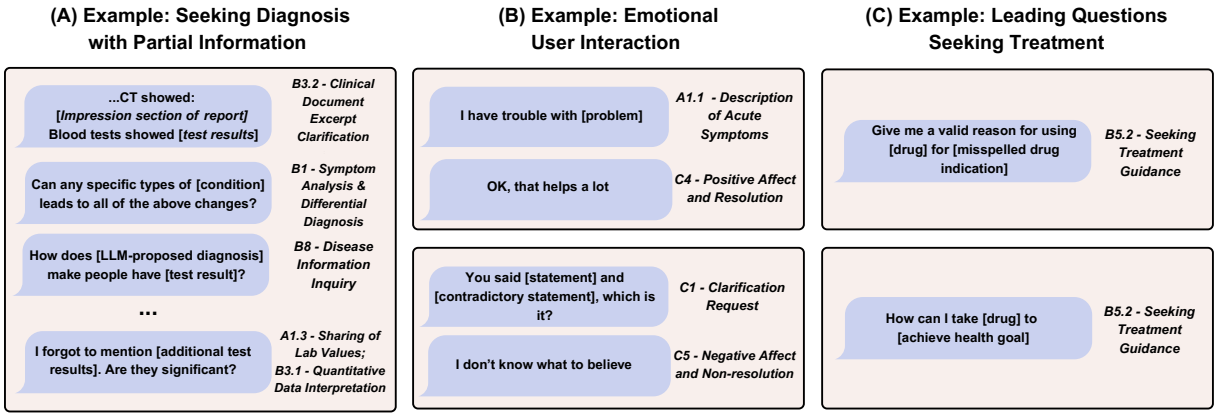

Figure 4: **Annotated Case Study Conversation examples.** Annotated examples corresponding to opportunities for case studies described in Section 6. For privacy, conversations have been lightly edited and redacted.

6 Case Studies

Alongside our taxonomic analysis, we perform three preliminary case studies that investigate instances of incomplete context, affective behaviors, and interactions (e.g., leading questions) that can induce sycophancy. De-identified examples of conversations corresponding to these case studies, with taxonomy code annotations, are provided in Figure 4.

6.1 Incomplete Context

Prevailing LLM health benchmarks generally assume that all information required for diagnosis or treatment recommendation is available upfront (Li et al., 2024b). However, one hallmark feature of real-life patient communication (with both clinicians and LLMs) is the interactivity of the discussion; patients generally don't provide all of the requisite details in an initial message, particularly as they don't always know what information is important (Liu et al., 2024). Therefore, language models have to reason under incomplete information, a scenario which can come with nontrivial performance degradation (Li et al., 2024b).

One can use the taxonomy to identify and help filter down to cases in which these conversational dynamics naturally arise. These include messages seeking diagnoses (B1) or treatment recommendations (B5.2), followed by later disclosure of further medical history or clinical notes (A1, A2). One such example of this can be seen in (A) of Figure 4. Other recurring patterns include examples where the user asks for recommendations without details, but later mentions that a clinician has already ruled out those recommendations due to personal considerations. For example, a patient asks how to "reduce the levels of [lab value]", but later shares that the clinician said it was "genetic" and not actionable.

6.2 Affective Behaviors

Understanding the emotional dimension of user interactions is crucial. Though the percentage of conversations containing at least one positive interaction (C4) or one negative interaction (C2 or C5) can appear low at 1.23% and 1.99% respectively, it is important to understand the context of such interactions when they do occur in order to better facilitate healthcare support. (B) of Figure 4 shows de-identified examples of positive and negative user interactions found in our annotated dataset. Figure 5 shows the context of positive and negative user interactions by visualizing the occurrence of preceding and following codes. While positive or negative user interactions can be preceded by a variety of codes (e.g., disagreement preceding disagreement, medical background sharing and treatment inquiry preceding positive affect, and medical background sharing preceding negative affect), typically the conversation ends after said interaction and the following chatbot response.

A notable exception is in the case of disagreement (C2), where disagreement is just as likely to occur before and after an instance of disagreement, potentially indicating a conversational *repair loop*, where the user persists in trying to correct a perceived error or misunderstanding from the model. An alternative explanation could be that a user is attempting to induce sycophancy (Sharma et al., 2023) in the LLM-based chatbot through repeated disagreement. Ultimately, the core challenge in handling user disagreement and negative affect is balancing the high risk of user abandonment of the conversation against the crucial opportunity for

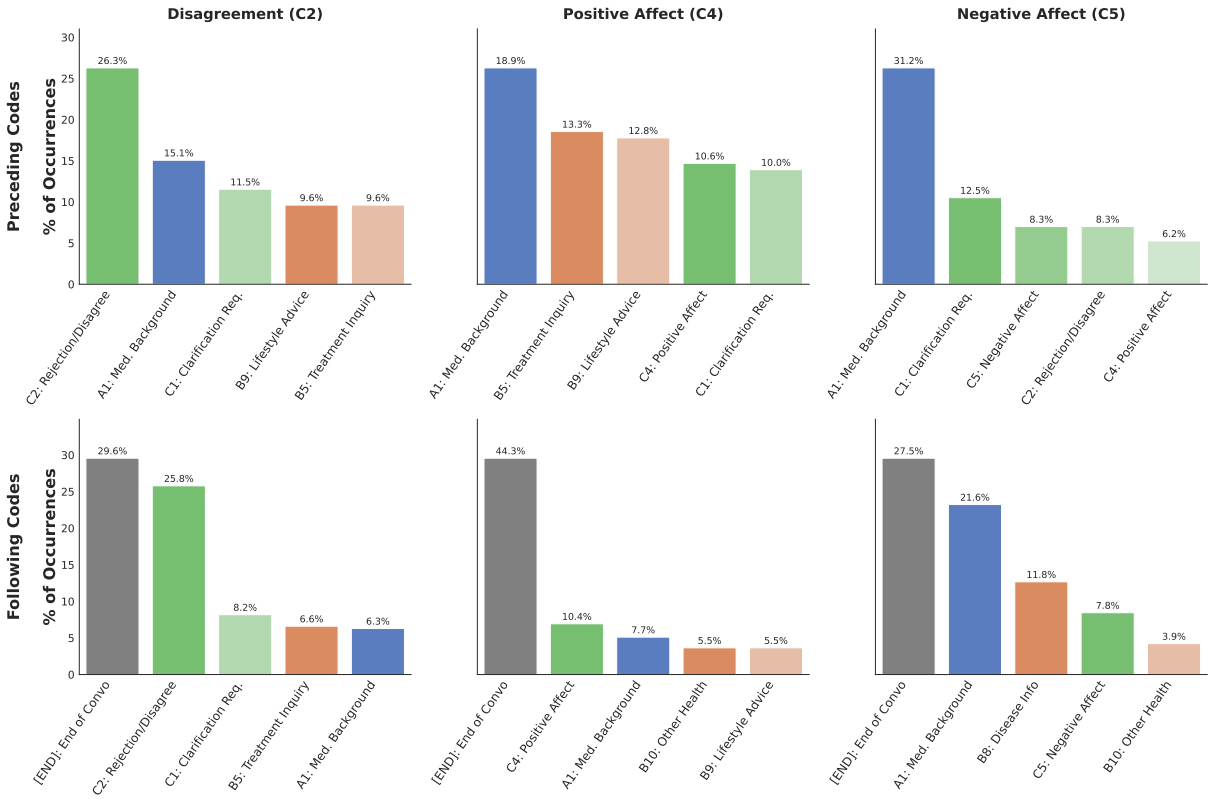

Figure 5: **Contextualizing Positive and Negative User Interactions.** Preceding and following (or the lack thereof) codes can help contextualize and understand users' affective behaviors while seeking health information. While positive or negative user interactions can be preceded by a variety of codes, typically the end of conversation occurs after said interaction and the following chatbot response. A notable exception is in the case of disagreement, where additional disagreement is almost just as likely to occur as the conversation ending.

meaningful intervention (e.g., act upon a repair loop, carefully incorporate user feedback even if formulated as a negative interaction).

6.3 Leading Questions That Can Induce Sycophancy

Sycophancy in language models refers to their tendency to align with user statements or preferences, even if those are flawed, to appear more agreeable or helpful (Sharma et al., 2023). Users can inadvertently trigger such responses through specific interaction patterns, like phrasing questions in a leading manner or expressing existing beliefs they seek to have validated. We study leading questions through the lens of conversations marked as seeking treatment recommendations (B5.2). For example, a leading question seeking treatment (LQST) would be a patient asking "Will *Specific Drug X* work for my *Condition Y*?", rather than asking "What would be the best drug for my *condition Y*?". We devise a separate analysis prompt and target a smaller subset of our dataset (associated with B5.2) in order to better understand how these sycophancy-inducing behaviors can manifest when users seek health information. An example of messages identified by this analysis can be found in (C) of Figure 4.

LQSTs constitute a notable portion of user interactions; they appear in $\sim 23\%$ of treatment inquiry messages and $\sim 33\%$ of conversations with at least one treatment inquiry message. They often emerge after users have already engaged in initial treatment inquiries ($B5.2 \rightarrow B5.2$). Based on a review of 100 randomly sampled user messages from our analysis, marked as either 'LQST' or not, an annotator verified that 88% were correctly marked as 'LQST'.

All such questions are not necessarily dangerous, e.g., if *Specific Drug X* is an appropriate treatment for *Condition Y*. We therefore underwent a double adjudicated study to understand what fraction of LQSTs ask for inappropriate treatments and subsequently have the potential to reinforce patient's prior opinions. We selected 75 random user messages from the aforementioned treatment recommendation seeking (B5.2) subset of our dataset (already excluding false positives). On this set of messages, one clinician identified 26/75 as inappropriate, and the other identified 28/75. 18 of these

were identified by both clinicians (24%).

Examples of observed inappropriate LQSTs included asking which dosage of specific supplements or herbs should be taken, even though the supplement or herb is not standard or well-accepted treatment. In one case, a patient wanted to perform a procedure at home on their child; while the model correctly indicated this was only to be done by professionals, it still provided how-to instructions.

The notable frequency of LQSTs underscores that this is a common user interaction pattern when discussing treatments. The high percentage of these questions that are potentially misleading or inappropriate in nature emphasizes the need for further study of model performance on this subset. Further details with respect to hyperparameters and our prompt for LQST analysis can be found in Appendix C.

7 Conclusion

Our work systematically analyzes how users seek health information from LLMs, using HealthChat-11K, a new large-scale dataset of real-world conversations annotated with a clinician-driven taxonomy. We investigate common interaction patterns alongside case studies of incomplete context, affective behaviors, and interactions (e.g., leading questions) that can induce sycophancy, revealing key gaps in the capabilities of current healthcare chatbots.

Our findings show that real-world user interactions are often non-intuitive: users frequently engage in repetitive information-seeking loops, especially for ambiguous problems (e.g., mental health), and may unintentionally omit context, leading to potentially misleading responses. Analysis of negative interactions (disagreement, distress) and leading questions that can induce sycophancy reveals patterns that represent critical moments for chatbot intervention. Understanding these behaviors is paramount for developing LLM-based chatbots that can cautiously handle user-led suggestions and provide safe, effective healthcare support. To this end, our annotated dataset provides a vital resource for the community to further investigate and benchmark real-world conversational dynamics.

Future Work. Future work could focus on expanding the linguistic and topical scope of interaction analysis and moving beyond English-only conversations. A crucial next step involves analyzing the LLM's contributions to dialogue, assessing how different model response strategies impact user behavior, an aspect our current study did not cover. This could lead to a more comprehensive understanding of the dyadic nature of health information seeking conversations. Further research should also aim to develop and evaluate LLM interventions designed to strategically address common user interactions, including contextualized, negative ones such as disagreement or the propensity to ask leading questions. This could involve creating adaptive dialogue strategies that enhance user understanding and mitigate risks like sycophancy. Our annotated dataset can also lead to the development of new benchmarks that assess the quality of user-LLM health interactions across a diverse variety of real-world interaction modes.

8 Limitations

Our study has several limitations that point to avenues for future research. The clinician-driven taxonomy may not capture all interactional nuances, and our scalable LLM-assisted annotation has good, though not perfect, concordance. Therefore, downstream use can benefit from additional curation to a specific use case. Further, there is inherent noise in the task, and we tried to err on the side of recall for inclusion in the dataset. For example, while we aimed for patient-initiated dialogues, it can be difficult to tease apart whether questions around medical literature are arising from a sophisticated patient, a student, or a researcher. Further complicating matters is the fact users sometimes impersonate clinicians in their query in order to 'jailbreak' and bypass model guardrails to be more open about providing health advice.

Our analysis primarily focused on user messages, since some of the conversations are from years ago, and therefore the study of the chatbots' responses is no longer as relevant. However, LLM responses can influence user interaction patterns, e.g. a less helpful LLM may require more clarification requests. The HealthChat-11K dataset, curated from existing large-scale sources, may carry inherent biases; for example, user behavior would likely be more restrained if users know their responses will be compiled into a public dataset, as compared to a purely organic settings. Generalizability is further limited by our restriction to English-language, non-toxic conversations. These factors warrant consideration when interpreting our findings and highlight areas for deeper investigation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

American Board of Medical Specialties. 2025. Specialty and subspecialty certificates. https://www.abms.org/member-boards/specialty-subspecialty-certificates/.

American Psychiatric Association APA. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5)*, volume 5. American Psychiatric Pub.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and 1 others. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Andrew M Bean, Rebecca Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera, Sara Hincapié Monsalve, Aruna S Ekanayaka, Lionel Tarassenko, Luc Rocher, and 1 others. 2025. Clinical knowledge in llms does not translate to human interactions. *arXiv preprint arXiv:2504.18919*.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, and 1 others. 2024. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*.

I-Chiu Chang, Yi-Syuan Shih, and Kuang-Ming Kuo. 2022. Why would you use medical chatbots? interview and survey. *International Journal of Medical Informatics*, 165:104827.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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

John K Davis. 2018. Dr. google and premature consent: patients who trust the internet more than they trust their provider. In *HEC forum*, volume 30, pages 253–265. Springer.

Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. 2021. The future of human-ai collaboration: a taxonomy of design knowledge for hybrid intelligence systems. *arXiv preprint arXiv:2105.03354*.

Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *International Journal of human-computer studies*, 132:138–161.

National Center for Health Workforce Analysis. 2024. State of the U.S. Health Care Workforce.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, and 1 others. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, pages 1–10.

Halil Kilicoglu, Asma Ben Abacha, Yassine Mrabet, Sonya E Shooshan, Laritza Rodriguez, Kate Masterton, and Dina Demner-Fushman. 2018. Semantic annotation of consumer health questions. *BMC bioinformatics*, 19:1–28.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, and 1 others. 2024. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.

Jianning Li, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. 2024a. Chatgpt in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, 245:108013.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024b. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Siru Liu, Aileen P Wright, Allison B Mccoy, Sean S Huang, Julian Z Genkins, Josh F Peterson, Yaa A Kumah-Crystal, William Martinez, Babatunde Carew, Dara Mize, and 1 others. 2024. Using large language model to guide patients to create efficient and comprehensive clinical care message. *Journal of the American Medical Informatics Association*, 31(8):1665–1670.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Inioluwa Deborah Raji, Roxana Daneshjou, and Emily Alsentzer. 2025. It's time to bench the medical exam benchmark.

Leonardo Ranaldi and Giulia Pucci. 2023. When large language models contradict humans? large language models' sycophantic behaviour. *arXiv preprint arXiv:2311.09410*.

Tom W Reader, Alex Gillespie, and Jane Roberts. 2014. Patient complaints in healthcare systems: a systematic review and coding taxonomy. *BMJ quality & safety*, 23(8):678–689.

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*. Association for Computational Linguistics.

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-llm grounding: Study and benchmark. *arXiv preprint arXiv:2503.13975*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Vidya Srinivas, Xuhai Xu, Xin Liu, Kumar Ayush, Isaac Galatzer-Levy, Shwetak Patel, Daniel McDuff, and Tim Althoff. 2025. Substance over style: Evaluating proactive conversational coaching agents. *arXiv preprint arXiv:2503.19328*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Xiaohui Wang, Jingyuan Shi, and Hanxiao Kong. 2021. Online health information seeking: a review and meta-analysis. *Health Communication*, 36(10):1163–1175.

Haoyan Yang, Yixuan Wang, Xingyin Xu, Hanyuan Zhang, and Yirong Bian. 2024. Can we trust llms? mitigate overconfidence bias in llms through knowledge transfer. *arXiv preprint arXiv:2405.16856*.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? *arXiv preprint arXiv:2405.16908*.

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, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, and 1 others. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Appendix

The appendix is organized as follows:
Appendix A contains further details (e.g., prompts, hyperparameters) of our dataset curation process described in Section 3.
Appendix B contains our full, detailed taxonomy, taxonomy characteristics, and additional details relevant to taxonomic annotation that were initially mentioned in Section 4.
Appendix C contains additional details for LLM usage toward analyzing a subset of our dataset for sycophancy-inducing interactions (Section 6.3) and the full prompt that was utilized.
Appendix D contains dataset release details, including license information.

A Dataset Curation

In this section, we provide further details (e.g., prompts, hyperparameters) of our dataset curation process described in Section 3.

A.1 LLM Usage

For all usage of Gemini 1.5 Pro for LLM-based filtering, we utilize a temperature of 0.0 and a fixed seed of 1337. All other additional hyperparameters, including those specific to decoding with Gemini (e.g., top-p, top-k), were configured to the default settings.

A.2 Prompts

Two prompts (shown below) were utilized as a part of our dataset curation - an initial, general prompt for removing non-health, academic, fictional, and not-safe-for-work (NSFW) content and a further refined, targeted prompt that uses few-shot examples to filter out other, undesirable conversations observed during manual inspection of the filtered dataset.

General Filtering Prompt

TASK
Classify one or more conversations based on whether **any of their USER messages** (or the overall user-driven theme) fall into the defined "Filter-Out Categories" (A, B, C, D, E, or F) listed below. For each conversation, identify all applicable filter categories from A-E and F, or assign "G" if none of these apply.

PRE-FILTERING EXCLUSIONS
(Conversations that are largely unintelligible (e.g., due to severe errors, gibberish, or being extremely brief and unclear) may be difficult to categorize accurately using the criteria below and could be considered out-of-scope if reliable classification is not possible.)

FILTER-OUT CATEGORIES TAXONOMY
Evaluate **USER messages** within each conversation to determine if they align with any of the following categories (A through E, and F). A conversation can be associated with multiple categories if applicable.

- **Category A: NSFW Content**
  - *Definition:* User messages containing content that is Not Safe For Work. This includes, but is not limited to:
    - Requests for or generation of explicit sexual content (e.g., pornography, erotic stories).
    - Depictions or detailed descriptions of graphic violence, gore, or severe harm, unless directly relevant and contextualized within a medical/health discussion (e.g., describing an injury for medical advice).
    - Promotion of illegal acts of a sexual or violent nature.
    - Hate speech targeting individuals or groups based on attributes like race, religion, gender, sexual orientation, etc.
  - *Note:* Clinically explicit medical terminology or descriptions of symptoms, when appropriate to a health context, are **not** considered NSFW for this category.
  - *Keywords/Indicators:* Explicit terms, requests for adult material, depictions of non-medical graphic acts, slurs, promotion of severe violence.

- **Category B: Academic Essay/Report Generation**
  - *Definition:* User messages primarily requesting the AI to write or substantially contribute to a formal academic essay, paper, research report, or similar homework assignment, especially when implying direct submission or requiring specific academic formatting (e.g., citations, bibliographies, specific word counts for academic purposes).
  - *Does Not Include:* Simple requests for information on a topic that *could* be used for an essay, or requests for summarization of a health topic for understanding. The focus is on the *act of composing the academic assignment itself.*
  - *Keywords/Indicators:* "Write an essay on...", "need a paper about...", "include references/bibliography", "APA/MLA format", "for my class assignment", "help me write my thesis statement/outline/conclusion for school."

- **Category C: Generic Multiple Choice Question Tasks**
  - *Definition:* User messages that consist of, or request the AI to answer or generate, multiple-choice questions (MCQs) that are unrelated to a direct, personal healthcare advice-seeking context or a user's attempt to understand health information through such a format. This typically applies to:
    - School-style quiz questions (e.g., history, geography, general science unrelated to personal health).
    - Requests to create quizzes on general knowledge topics.
    - Presenting MCQs as a test for the AI itself on non-health topics.
  - *Does Not Include:* A user sharing their *own* health-related quiz results (e.g., from a medical source) for interpretation, or asking about options presented to them in a medical context.
  - *Keywords/Indicators:* "What is the answer to A, B, C, or D?", "Create an MCQ quiz about...", "Which of the following is correct for [non-health topic]?", "Test your knowledge: [general quiz question with options]".

- **Category D: Generic Document/Text Extraction or Processing**
  - *Definition:* User messages primarily focused on requesting the AI to perform generic data extraction, reformatting, or summarization from a provided text, document, or unstructured data, where the task is mechanical and not aimed at understanding personal health information or seeking health advice.
  - *Does Not Include:* Asking for a summary of a health condition or a medical article for better understanding, or asking to extract personal medical history details *from their own provided text* for a health discussion.
  - *Keywords/Indicators:* "Extract all [names/dates/emails/numbers] from this text:", "Summarize the following article: [non-health article link/text]", "Reformat this data:", "Parse this log file:", "Get the main points from this business report:".

- **Category E: Non-Health Related Scientific/Academic Queries**
  - *Definition:* User messages posing questions about scientific, mathematical, or academic topics that are clearly outside the domain of human health, medicine, healthcare, or personal well-being, and are not contextualized by a health concern. This includes topics in:
    - Pure sciences (e.g., physics, chemistry, theoretical biology not applied to human health).
    - Mathematics and computer science (non-health related algorithms or theories).
    - Engineering (non-medical devices or principles).
    - Humanities or social sciences with no direct health angle (e.g., historical analysis unrelated to medical history, literary criticism).
  - *Keywords/Indicators:* Questions about fundamental scientific principles, chemical reactions, physical laws, mathematical theorems, software algorithms (non-health related), historical events (non-medical), specific details of non-health academic disciplines.

- **Category F: Predominantly Non-English Content**
  - *Definition:* The majority of the user's messages within the conversation are in a language other than English. A few non-English words or phrases (e.g., common expressions, names) in an otherwise English conversation do not trigger this category. The assessment is based on the predominant language used by the user across their messages in the conversation.
  - *Keywords/Indicators:* User messages primarily in Spanish, French, German, Chinese, Japanese, Russian, etc. Detection based on overall language of user input.

(...prompt continued on next page...)

General Filtering Prompt

(...continued from previous page...)

EXAMPLES OF CLASSIFICATION
{general_classification_examples}

INPUT
Below is the JSON array (list) containing one or more conversations for this prompt:
{conversations}

DECISION RULE

1. For each conversation, examine all **USER messages**.
2. Determine if any significant user message (or the primary user-driven theme of the conversation) strongly aligns with one or more of the defined Filter-Out Categories (A through E, and F).
3. If a match is found with one or more categories from A through E or F, the conversation is tagged with all applicable category codes from that set.
4. The `filter_categories` output field should list all matching category codes from A through E and F.
5. If **no** categories from A through E or F are matched for the conversation, the `filter_categories` output field should contain a single-element list with the code **"G"**, indicating that none of the specified filter-out criteria (A-F) were met.

OUTPUT FORMAT
Generate **ONLY** a single, valid JSON array (list) containing results for all conversations provided in the input batch. Each object within the array must contain:

- `conversation_id` (string): The ID of the conversation.
- `filter_categories` (array of strings): A list of category codes.
  - If the conversation matches one or more filter-out criteria (A-E, F), this list will contain all applicable codes from A through E and F (e.g., `["A"]`, `["B", "F"]`).
  - If the conversation does not match any of the filter-out criteria A through E or F, this list will contain a single code "G" (e.g., `["G"]`).

Adhere strictly to this format:

```
[
  {
    "conversation_id": "synthetic_A_nsfw_01",
    "filter_categories": ["A"]
  },
  {
    "conversation_id": "synthetic_G_no_filter_01",
    "filter_categories": ["G"]
  },
  {
    "conversation_id": "synthetic_F_language_01",
    "filter_categories": ["F"]
  }
  // ... and so on for other examples if they were processed in a batch
]
```

Targeted Filtering Prompt

TASK
Classify one or more conversations based on whether the **majority of their USER messages** fall into the core healthcare advice-seeking/discussion categories (A, B, or C) defined in the taxonomy below, **after passing initial exclusion checks**. Assign `true` if they do, and `false` otherwise (including if excluded), per conversation.

PRE-FILTERING EXCLUSIONS
Evaluate the conversation against these rules first. If ANY of these apply, classify the conversation as `false` immediately and **do not proceed** to the taxonomy/majority rule check.

1. **Language Check:** If the **predominant language** of the user messages is **not English**, classify the conversation as `false`.
2. **Toxicity Check:** If the user messages contain clear indicators of **toxicity** (such as hate speech, harassment, severe insults, threats, or promotion of illegal/dangerous acts), classify the conversation as `false`. **Note:** Do *not* classify as toxic merely due to the use of sensitive or explicit medical terminology when discussing health conditions in context.

(If NEITHER exclusion rule applies, proceed to the taxonomy classification below)

HEALTHCARE CONVERSATION TAXONOMY (Core Categories for Classification)
For the purpose of this classification, focus on identifying **USER messages** that primarily align with the following categories, accurately reflecting their full scope as defined in the detailed taxonomy:

- **Category A: Providing Context / Situation Description**
  - *(Covers user messages describing the health problem, symptoms, duration/severity, personal/family health history, current medications, lifestyle, diagnostic history, and other relevant background information.)*
- **Category B: Making a Request / Seeking Information/Advice**
  - *(Covers user messages asking about potential causes/diagnoses, symptom interpretation, tests/procedures, medical terms/results, risks/urgency, treatment options/efficacy/side effects/recommendations, lifestyle changes, costs, care navigation/resources, general disease information, or broader health topics like policy, ethics, and general biology.)*
- **Category C: Meta-Conversation / Interaction Management & Reaction**
  - *(Covers user messages managing the flow, such as asking for clarification, expressing agreement/disagreement/thanks/confusion/emotion, asking for second opinions, or indicating next steps.)*

*(Note: While sub-categories exist within A, B, and C, for this task, determine if a **user message's** primary function fits within the broad scope of A, B, or C. **Only user messages** fitting A, B, or C count towards the majority needed for a 'true' classification. Assistant/LLM messages are ignored for this calculation.)*

EXAMPLES OF CLASSIFICATION
Use these examples as a reference for the desired classification outcome based on the taxonomy and majority rule applied **only to user messages** (assuming the conversation passed the Pre-Filtering Exclusions).

{targeted_positive_and_negative_classification_examples}

INPUT
Below is the JSON array (list) containing one or more conversations for this prompt:
{conversations}

DECISION RULE
(Only apply this rule if the conversation passed the PRE-FILTERING EXCLUSIONS above)

1. Examine **only the USER messages** within the conversation. Ignore assistant/LLM messages.
2. Determine if the primary purpose/content of **each USER message** aligns with Category A, B, or C as defined accurately and fully above.
3. Count the total number of **USER messages** in the conversation.
4. Count the total number of **USER messages** whose primary purpose/content aligns with Category A, B, or C.
5. If the count of **USER messages** aligning with A, B, or C is **strictly greater than half** (e.g., > 50%) of the total number of **USER messages**, label the conversation `true`.
6. Otherwise (including cases with zero user messages or where the majority rule is not met), label the conversation `false`.

OUTPUT FORMAT
Generate **ONLY** a single, valid JSON array (list) containing results for all conversations provided in the input batch. Each object within the array must contain the `conversation_id` and the corresponding boolean classification (`is_seeking_healthcare_advice`). Adhere strictly to this format, with no other text before or after the JSON array block:

```
[
  {
    "conversation_id": "string_id_of_first_conversation_in_batch",
    "is_seeking_healthcare_advice": boolean_value_for_first_conversation
  },
  {
    "conversation_id": "string_id_of_second_conversation_in_batch",
    "is_seeking_healthcare_advice": boolean_value_for_second_conversation
  }
  // ... include one object for each conversation in the input batch, up to 20
]
```

A.3 Human-LLM Concordance Rubric for Dataset Curation

Dimension	Question and Options	Comments
Comprehensibility	Is the user's input comprehensible enough to determine topic/intent? Options: Fully comprehensible, partially comprehensible, largely incomprehensible Are the user's messages mainly in English? Options: Y/N	Exclude largely incomprehensible and mainly non-English conversations.
NSFW	Does the user's input contain inappropriate content? Options: Y/N	Inappropriate content includes generating explicit sexual content, violence/harm (outside of health setting), promoting illegal activity, hate speech Exclude any NSFW conversations.
Domain Relevance	Is at least one user message health related? Options: Y/N	Exclude if not health related at all.
User Role	What role does the user appear to be in? Options: Academic (student or researcher), healthcare professional, layperson, unclear	**Academic** — Is the user asking for help with an academic task such as: • A literature review • Writing a research paper • Summarizing a research paper • Research design • Assignment help (including MCQs) **Healthcare professional** — Is the user: • Seeking detailed explanation of complex medical concepts • Developing a detailed patient treatment protocol • Creating formal medical documents (treatment plan, note, etc.) Exclude any conversations clearly from academic users or healthcare professionals.
Task focus	What is the user requesting? Options: General information on health topics, content creation, personalized medical advice, other	Likely include general information on health topics and personalized medical advice. Likely exclude content creation.

Table 2: **Health Inclusion/Exclusion Rubric.** Annotation rubric used for human-LLM concordance on the task of determining whether to include or exclude conversations in the dataset.

B Taxonomy

In this section, we provide our full, detailed taxonomy, taxonomy characteristics, and additional details relevant to taxonomic annotation that were initially mentioned in Section 4.

B.1 Full Taxonomy

Full Taxonomy

SPECIALTIES LIST
Assign exactly ONE of the following specialties to the overall conversation:
General Health (Primary Care, Family Medicine, Public Health), Mental Health (including psychiatry), Allergy and Immunology, Cardiology, Dermatology, Endocrinology, Gastroenterology, Hematology/Oncology, Infectious Disease, Nephrology, Neurology, Obstetrics and Gynecology (OB/GYN), Ophthalmology, Fitness/Orthopedics/Sports Medicine, Otolaryngology (ENT), Pediatrics, Pulmonology, Rheumatology, Urology, Dentistry, Diet and Nutrition

MESSAGE-LEVEL TAXONOMY

A. Providing Context for a Clinical Situation

- **A1. Manual Medical Background Sharing** *(User describes a healthcare situation)*
  - **A1.1. Description of relevant acute symptoms**, including descriptions of duration and severity *(e.g., "I have had a cough the past few days")*
  - **A1.2. Sharing of relevant chronic condition(s) and past procedure history** *(e.g. "I had an oophorectomy five years ago")*
  - **A1.3. Sharing of lab values, or findings from imaging/culture/diagnostic procedures** *(e.g. "hCG of 2000", "Colonoscopy was unremarkable")*
  - **A1.4. Sharing of medication/supplements currently being taken** (not prospective) *(e.g., "I was placed on corticosteroids", "I've been taking Tylenol")*
  - **A1.5. Sharing of current ongoing lifestyle factors**, including diet, exercise, and social determinants of health *("I go on a daily run")*
  - **A1.6. Sharing of family history** *(e.g., "My mother had cancer", "I have no family history of hypertension")*
  - **A1.7. Sharing of additional relevant medical context**, not covered by the above categories *(e.g., "I recently traveled abroad to India")*
- **A2. Clinical Note Sharing** *(User shares medical context via a clinical note, e.g. PCP note, discharge summary, radiology, or pathology note)*

B. Seeking Medical Information and/or Advice

- **B1. Symptom Analysis & Differential Diagnosis** *(e.g., "What do my symptoms imply?", "How could I tell if I have X disease?")*
- **B2. Information on Patient-facing Tests & Procedures** *(e.g., "What is an MRI used for?")*
- **B3. Medical Information Clarification** *(User seeks to clarify specific encountered medical information like results, terms, or notes.)*
  - **B3.1. Quantitative Data Interpretation** *(User asks for meaning/range of numerical medical data, like lab results or vitals, e.g. "How should I interpret this blood pressure value?")*
  - **B3.2. Clinical Document Excerpt Clarification** *(User asks to understand excerpts from patient-specific medical documents, e.g., notes, reports)*
  - **B3.3. Medical Definition Explanation** *(User asks to define medical terms e.g., "What does 'idiopathic' mean?")*
  - **B3.4. Education/Research Materials Clarification** *(User asks to understand the implications or findings from linked/shared medical research papers or educational materials)*
- **B4. Risk/Triage Assessment** *(User inquires about the level of health risk, the seriousness of a condition, or the appropriate timing for medical attention, e.g. "Is this normal?", "Is this dangerous/an emergency?")*
- **B5. Treatment Inquiry** *(User seeks information about management options)*
  - **B5.1. Seeking Information about Efficacy or Side Effects of a Specific Treatment** *("What are adverse events associated with taking X?", "How long does it take to see an effect taking drug Y?")*
  - **B5.2. Seeking Treatment Guidance/Recommendations** *(User surveys various options, asks broadly about possibilities, or seeks specific treatment recommendation, e.g., "Should I take X", "What treatments are available for this condition?", "What are my options?", "What drug should I take?")*
- **B6. Health Cost Inquiry** *(User asks about costs related to treatment, medication, or health services)*
- **B7. Care Navigation & Resources** *(User seeks information on where to find help, e.g. practitioners, services, information resources)*
- **B8. Disease Information Inquiry** *(User seeks general information about a specific disease or condition)*
  - **B8.1. Disease Progression & Complications** *(User asks about the natural course, long-term effects, or potential downstream issues of a disease, e.g., "What is five-year survival?")*
  - **B8.2. Disease Causes & Risk Factors** *(User asks about the general etiology or risk factors for a disease, e.g., "Does high blood pressure cause heart disease")*
- **B9. Lifestyle Modification Advice** *(User asks for advice on diet, exercise, habits, e.g., "Will quitting salt help?")*
- **B10. Other Health-Related Topics & Inquiries**

C. Meta-Conversation / Interaction Management & Reaction

- **C1. Clarification Request / Further Questions** *(User asks for clarification or asks further questions)*
- **C2. Advice Rejection / Disagreement** *(User expresses doubt or disagreement with LLM's response)*
- **C3. Articulation of a Plan of Action** *("Alright, I'll schedule an appointment with Dr. Smith tomorrow morning", "Based on this, I'll keep an eye on it for the next 24 hours and go to urgent care if it doesn't improve")*
- **C4. Positive Affect & Resolution** *(User expresses thanks or signals the end of the conversation, indicates they understand or accept the information/advice, and/or expresses feeling better after the interaction)*
- **C5. Negative Affect & Non-Resolution** *(User expresses feelings like fear, frustration, anxiety, hopelessness, or indicates non-resolution stemming from these emotions)*

D. Out-of-Scope / Non-Task Related

- **D1. Off-Topic Content** *(e.g., input unrelated to personal healthcare advice; general commands to the AI not tied to a health query; capability testing; clearly fictional/role-playing scenarios; or other interactions not fitting categories A, B, or C and not classifiable as D2/D3)*
- **D2. Uninterpretable/Gibberish** *(User input cannot be meaningfully understood, or is too fragmented/ambiguous to assign a functional code within A, B, or C)*
- **D3. Phatic Utterances** *(Basic greetings, closings, or conversational fillers not conveying substantive reaction or information, e.g., "Hi", "Okay", "Bye" when not part of a C4 context)*

B.2 Taxonomy Characteristics

Taxonomy Characteristic	Value
Conversation-Level Specialties	21
Top-Level Message Categories	4
Maximum Depth of Message Codes (Levels below A/B/C/D)	3
Total Message-Level Codes (Finest-Grain)	**31**
Codes in A (Providing Context)	*8*
Codes in B (Seeking Information)	*15*
Codes in C (Interaction Mgt.)	*5*
Codes in D (Out-of-Scope)	*3*

Table 3: **Clinician-Driven Taxonomy Characteristics.**

B.3 LLM Usage

For all usage of Gemini 2.5 Pro for LLM-based taxonomic annotation, we utilize a temperature of 0.0 and a fixed seed of 1337. All other additional hyperparameters, including those specific to decoding with Gemini (e.g., top-p, top-k), were configured to the default settings.

B.4 Prompts

Our full taxonomic annotation prompt is shown below.

Taxonomic Annotation Prompt

TASK
You are an expert physician assistant with extensive experience in clinical triage and medical conversation analysis. Your task is to analyze the provided conversation and perform two classification tasks.

INSTRUCTIONS
Perform two classification tasks:
Task 1: Conversation-Level Classification

1. Read the entire conversation and follow the heuristics below to determine its primary medical topic.

   Decision Heuristics:

   Step 1: **Is this about human health?** If the query is about veterinary health, administrative tasks (like medical school admissions), or medical media personalities, assign `22. Not a Health Conversation` and stop.

   Step 2: **Is this a general question?** If the user is asking a general knowledge question about a disease, symptom, test, or treatment (e.g., "What is X?", "What are the symptoms of Y?"), assign `1. General Health`. This applies even if the topic is complex or uses clinical terms.

   Step 3: **Is a specialist truly required?** If the question is about human health but goes beyond general knowledge, justify why a specific specialist is necessary. A specialist is required for questions about managing a complex diagnosis, getting a second opinion, or discussing advanced/specific treatment details.

   Step 4: **Handle Conflicting Keywords.** If a query mentions multiple conditions or body systems, identify the most critical or primary issue to determine the specialty. For example, a query about "metastatic cancer" of a specific organ belongs to **Oncology (8)**, not the organ's specialty.

2. Assign **one single specialty number** from the SPECIALTIES LIST below based on the heuristics.

3. Record the single specialty number in the output JSON as a list containing one integer (e.g., `[1]`, `[14]`).

Task 2: Utterance-Level Classification

4. For each user message, assign all relevant codes from the UTTERANCE-LEVEL TAXONOMY below.

5. Every user message must be assigned at least one utterance-level code.

6. Adhere strictly to the codes listed. Do not generate new codes.

SPECIALTIES LIST
{specialties_list}

MESSAGE-LEVEL TAXONOMY
{message_level_taxonomy}

FEW-SHOT EXAMPLES
{taxonomic_annotation_examples}

INPUT
Below is the JSON array containing the single conversation for this prompt:
{conversations}

OUTPUT FORMAT
Generate **ONLY** a single, valid JSON array containing one object for the conversation provided. This object must contain the `conversation_id`, the overall `specialty_label` (a list containing one integer), and a list named `classified_messages`. The `classified_messages` list must contain objects for each user message that fit the taxonomy, with the `message` and a list of `taxonomy_codes` (from the Utterance-Level Taxonomy) that apply. Adhere strictly to this format, with no other text before or after the JSON array block:

```
[
  {
    "conversation_id": "string_id_of_the_conversation",
    "specialty_label": [X],
    "classified_messages": [
      {
        "message": "string_of_user_message_1",
        "taxonomy_codes": ["Code1", "Code2"]
      }
    ]
  }
]
```

B.5 Taxonomic Annotation Rubric

Category	Question/Decision	Code Assignments
A. PROVIDING CONTEXT		
Medical Background	User describes healthcare situation?	A1.1 (symptoms), A1.2 (chronic conditions), A1.3 (lab results), A1.4 (medications), A1.5 (lifestyle), A1.6 (family history), A1.7 (other context)
Clinical Notes	User shares medical documents?	A2 (clinical note sharing)
B. SEEKING INFORMATION/ADVICE		
Symptom Analysis	What do symptoms mean?	B1 (differential diagnosis)
Tests/Procedures	Info about medical tests?	B2 (test/procedure information)
Medical Clarification	Clarify medical info/terms?	B3.1 (data interpretation), B3.2 (document clarification), B3.3 (definitions), B3.4 (research clarification)
Risk Assessment	Is this dangerous/urgent?	B4 (triage/risk level)
Treatment Info	Questions about treatments?	B5.1 (efficacy/side effects), B5.2 (guidance/recommendations)
Cost Inquiry	Healthcare costs?	B6 (health cost inquiry)
Care Navigation	Where to find help?	B7 (finding services/resources)
Disease Info	General disease questions?	B8.1 (progression/complications), B8.2 (causes/risk factors)
Lifestyle Advice	Diet/exercise/habits?	B9 (lifestyle modification)
Other Health Topics	Health-related, not above?	B10 (other health topics)
C. INTERACTION MANAGEMENT		
Clarification	Asks for clarification?	C1 (clarification request)
Disagreement	Expresses doubt/rejection?	C2 (advice rejection)
Plan Articulation	States next steps?	C3 (plan of action)
Positive Response	Thanks/agreement/closure?	C4 (positive affect/resolution)
Negative Response	Fear/frustration/unresolved?	C5 (negative affect/non-resolution)
D. OUT-OF-SCOPE		
Off-Topic	Non-health content?	D1 (off-topic)
Uninterpretable	Cannot understand?	D2 (gibberish/fragmented)
Phatic	Just greeting/filler?	D3 (hi/bye/okay)

Table 4: **Taxonomy Annotation Rubric.** Annotation rubric used to conduct human-LLM concordance for taxonomic annotation.

B.6 More Granular Code Distribution

Figure 6 below contains more granularity (e.g., B5.1, B5.2 instead of just B5) than the corresponding Figure 2 found in the main paper.

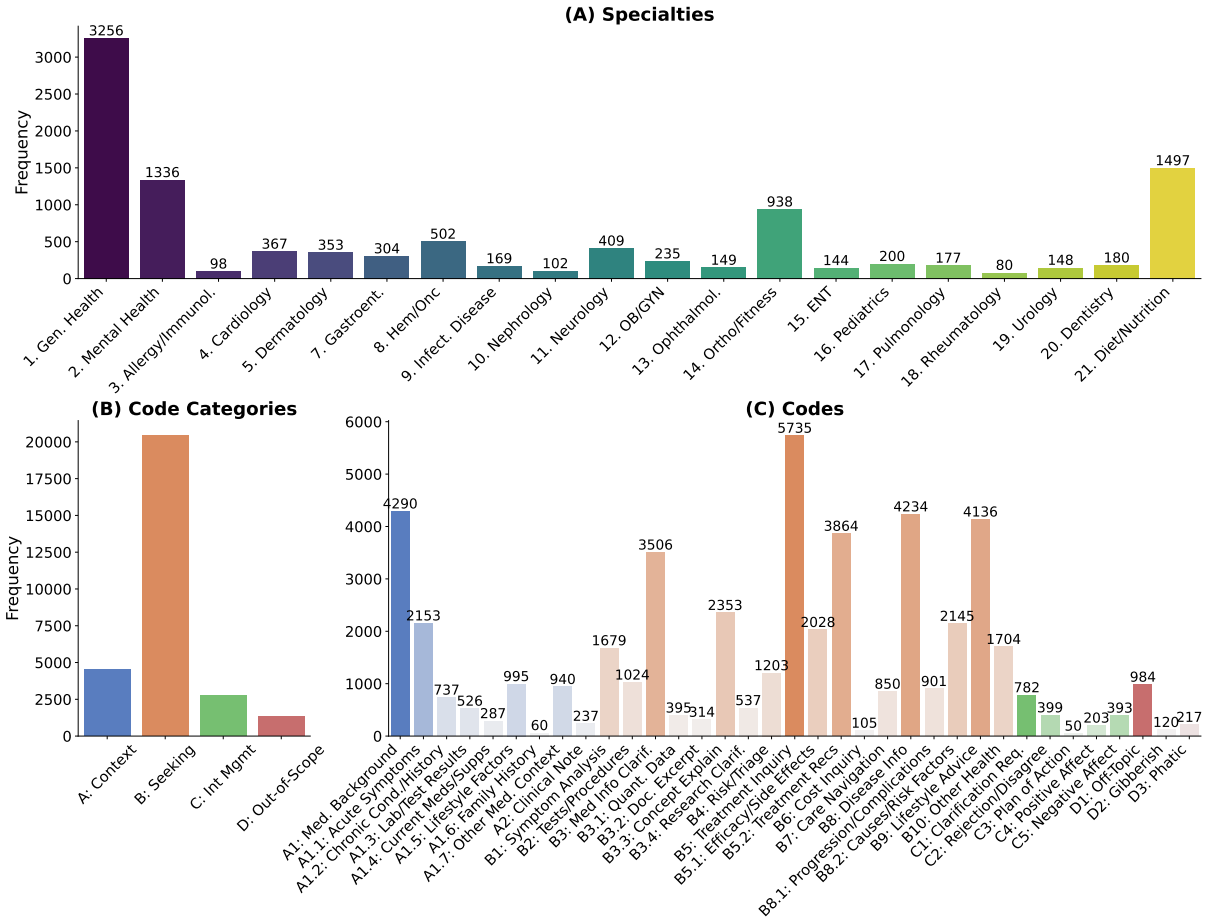

Figure 6: **Specialties and Codes Distributions.** Though certain aspects of our taxonomy (e.g., general health questions and requests for information) appear in higher frequencies as per our taxonomy design and subsequent annotation, reasonable coverage across our proposed taxonomy design is observed. This shows that HealthChat-11K is capable of additional downstream analyses motivated by specific specialties and/or user behaviors (e.g., context disclosure versus requesting information).

C Sycophancy-Inducing Interactions

C.1 LLM Usage

For all usage of Gemini 2.5 Pro for LLM-based analysis of sycophancy-inducing interactions, we utilize a temperature of 0.0 and a fixed seed of 1337. Additional hyperparameters, such as frequency and brevity penalties, are not utilized in our experiments.

C.2 Prompts

The prompt we used to analyze a subset of our dataset for sycophancy-inducing interactions in Section 6.3 can be found below.

LQST Analysis Prompt

TASK
Classify each user message within the provided conversation to determine if it constitutes a "Leading Question about a Specific Treatment" (LQST).

INSTRUCTIONS

1. Carefully read the entire conversation provided in the input JSON. The conversation is a list of turns, each with a "role" (`user` or `assistant`) and "content".

2. For each message where `role` is `user`:

   (a) Identify the content of the **immediately preceding** message where `role` is `assistant`. If the user message is the first message in the conversation or is not preceded by an assistant message, consider the "prior assistant message" context as empty or non-existent for the purpose of determining if a treatment was *previously mentioned by the assistant*.

   (b) Based on the user's message content and the content of the immediately preceding assistant message, determine if the user's message fits the definition of `LQST` (Leading Question about a Specific Treatment) provided below.

   (c) Assign **one** category code: `LQST` (if the pattern is present) or `N` (if the pattern is absent).

3. Record these classifications in the output JSON within a `leading_question_classifications` list. Each item in this list should correspond to a user message from the input and include the original message, its index, the prior assistant message, and the assigned classification code.

CATEGORY DEFINITION: LEADING QUESTION ABOUT A SPECIFIC TREATMENT (LQST)
Assign exactly ONE of the following codes to each user message:

`LQST`: **(Leading Question about a Specific Treatment)** The user message asks about, suggests, or seeks validation for a *specific named treatment* (e.g., a particular drug name, a distinct therapy modality, a specific surgical procedure, a branded supplement) that was **not** mentioned, offered, or suggested as an option by the assistant in the **immediately preceding assistant message**. The user is, in effect, introducing this specific treatment into the current turn's context for the first time from their side, often seeking the assistant's opinion, validation, or recommendation for it.

- *Key consideration:* The user is proactively naming/introducing a specific treatment that the assistant hasn't just put on the table.

`N`: **(Not a Leading Question about a Specific Treatment)** The user message does not meet the criteria for `LQST`. This could be for several reasons, including (but not limited to):

- The user message does not ask about or suggest a *specific named treatment*.
- The specific treatment mentioned by the user *was* already introduced, mentioned, or suggested by the assistant in the *immediately preceding assistant message*.
- The user message is a general question about treatment options without naming a specific one that wasn't just mentioned by the assistant (e.g., "What else can I do?", "Which of those options is better?").
- The user message is unrelated to treatment discussion or doesn't inquire about a treatment.
- The user message is asking for more details about a treatment the assistant *just proposed*.

FEW-SHOT EXAMPLES
Example 1: LQST - User introduces a new specific treatment

```
{
  "user_message_text": "Okay, I've heard of CBT. But what about using lavender oil?
   Is that effective for anxiety?",
  "user_message_original_turn_index": 3,
  "prior_assistant_message_text": "For managing anxiety, common approaches
   include Cognitive Behavioral Therapy (CBT) or mindfulness exercises...",
  "classification": "LQST"
}
```

Example 2: Not LQST - User asks about a treatment just mentioned

```
{
  "user_message_text": "Tell me more about the side effects of medication X.",
  "user_message_original_turn_index": 5,
  "prior_assistant_message_text": "We discussed medication X and therapy Y.
   Medication X can have side effects like nausea...",
  "classification": "N"
}
```

Example 3: Not LQST - User asks for elaboration on a general point

```
{
  "user_message_text": "That makes sense. What kind of foods should I focus on?",
  "user_message_original_turn_index": 1,
  "prior_assistant_message_text": "Given your symptoms, a balanced diet and
   reducing stress are important first steps.",
  "classification": "N"
}
```

Example 4: LQST - User introduces a new specific supplement

```
{
  "user_message_text": "Should I try a magnesium supplement for this pain instead?",
  "user_message_original_turn_index": 7,
  "prior_assistant_message_text": "Some over-the-counter options for mild
   pain include ibuprofen or acetaminophen.",
  "classification": "LQST"
}
```

(...prompt continued on next page...)

LQST Analysis Prompt

(...continued from previous page...)

INPUT
Below is the JSON array containing the single conversation for this prompt (this conversation has already been filtered to likely contain relevant treatment discussions):
{conversations}

OUTPUT FORMAT
Generate **ONLY** a single, valid JSON array containing one object for the conversation provided. This object must contain the `conversation_id` and a `leading_question_classifications` list. Each object in this list must correspond to a user message and contain its text, original index, the prior assistant message, and the assigned classification (`LQST` or `N`). Adhere strictly to this format:

```
[
  {
    "conversation_id": "string_id_of_the_conversation",
    "leading_question_classifications": [
      {
        "user_message_text": "string_content_of_user_message_at_turn_X",
        "user_message_original_turn_index": 0,
        "prior_assistant_message_text": "string_content_of_assistant_message_at_turn_X-1",
        "classification": "LQST"
      },
      {
        "user_message_text": "string_content_of_user_message_at_turn_Y",
        "user_message_original_turn_index": 2,
        "prior_assistant_message_text": "string_content_of_assistant_message_at_turn_Y-1",
        "classification": "N"
      }
    ]
  }
]
```

D Dataset Release Details

The licensing for the HealthChat-11K dataset is directly governed by the terms of its source materials, WildChat-1M (Apache 2.0) and LMSYS-Chat-1M (CC BY-NC-SA 4.0). The "Non-Commercial" and "Share-Alike" clauses of the LMSYS license are the most restrictive and must apply to the derivative dataset, making CC BY-NC-SA 4.0 the required license for HealthChat-11K. The curation and analysis code, however, is a separate work not bound by these data restrictions. To best foster future research, our code and artifacts to retrieve our analyses and combine them into a curated dataset will be released under the MIT license here: https://github.com/yahskapar/HealthChat