HealthConvos: A physician-annotated dataset of presenting complaints sourced from real, multi-turn patient conversations

Tony Yue Sun[1], **Justin Yu**[1], **Angel Samsuddin Maredia**[1], **Jaisal Friedman**[1], **Uzair Khan**[1], **Sebastian Wakefield**[1], **Isabel Zaller**[1], **Daniel Bongiolatti**[1], **Jessica Fan**[1], **Amit Bhatt**[1], **David C. Whitehead**[1,2], **Rishi Khakhkhar**[1,3], **Muthuraman Alagappan**[1], **Cian Owen Hughes**[1]

[1] *Counsel Health, New York, NY*
[2] *Harvard Medical School, Boston, MA*
[3] *Icahn School of Medicine at Mount Sinai, New York, NY*

TONY@COUNSELHEALTH.COM

Abstract

We present *HealthConvos*, an anonymized dataset of multi-turn physician consultations based on real patient conversations from Counsel Health's asynchronous telehealth platform. In order to satisfy HIPAA Safe Harbor requirements, protected health identifiers were mapped to equivalent surrogates (e.g., patient names replaced by realistic pseudonyms), and thread messages were paraphrased using a few-shot prompted large language model. Each HealthConvos thread has been physician-labelled with clinical outcomes documentation (e.g., SOAP notes) to represent the disposition and management of ongoing patient concerns. Physicians also adjudicated the escalatory risk of threads to highlight cases where patients should seek out more immediate forms of in-person care. By releasing this dataset, researchers can evaluate history-taking chatbots against (1) the most relevant questions physicians asked given existent information within a thread, (2) the SOAP notes generated at particular turns within a conversation, and (3) the escalatory risk of a particular thread.

Keywords: Multi-turn conversational dataset, dataset release, evaluations and benchmarks

Data and Code Availability The de-identified dataset can be publicly downloaded from the HealthConvos Access Github after users agree to the license and data usage agreement.

Institutional Review Board (IRB) The HealthConvos dataset release did not require IRB approval; all data were de-identified per HIPAA Safe Harbor and shared with user consent.

1. Introduction

As large language models (LLMs) have improved, patients have increasingly turned to chatbot experiences to discuss their ongoing acute and chronic medical needs (McCaffery et al., 2024; Presiado et al., 2024). While foundational models are now significantly better at correctly answering questions on the American medical licensing exams (Saab et al., 2024; Wang et al., 2025; de Paiva et al., 2025; Abbas et al., 2024), their ability to accurately triage and provide relevant diagnostic advice to patients is still under active investigation (Masanneck et al., 2024; Chen et al., 2024). To aid in model benchmarking and evaluations, researchers have released a limited number of conversational health datasets; however, but most are purely synthetic examples due to the challenges and risks associated with collecting and de-identifying real patient conversations (Arora et al., 2025; Xu et al., 2025).

In this publication, we introduce *HealthConvos*, the first de-identified, multi-turn conversational dataset derived from real patient interactions across Counsel Health's asynchronous telehealth platform. In addition to the patient + AI system messages, the dataset includes physician responses (which includes the most relevant questions that physicians sought to ask given the existing information within the thread), de-identified physician annotations of patient outcomes (i.e., summary SOAP notes generated post-encounter), as well as an physician-labeled escalation risk for the thread. To ensure patient and provider privacy, all conversations and physician notes were de-identified by replacing protected health informa-

tion (PHI) with realistic "surrogate" candidates and paraphrased using a large language model (LLM).

2. Methods and Processing

We extracted 149 curated threads containing patient-initiated physician consultations from Counsel Health's asynchronous telehealth platform during Counsel's closed alpha testing period (July 25, 2025 through September 1, 2025). As some context, Counsel Health's platform utilizes an LLM-driven pre-visit intake process that initially connects patients with an agentic chatbot (referred to in the dataset as the "system" message sender). The agentic chatbot's primary role is to engage patients in history-taking (i.e., by asking contextually relevant follow-up questions, querying the patient's medical record, or retrieving medical literature and guidelines from both online and offline sources). Patients are free to invite physicians into their threads at any time (e.g., to order or refill medications, to receive more formal diagnoses after history-taking, or to coordinate follow-up care). This dataset only includes threads in which a physician was explicitly invited to the chat.

Patients sent slightly more total messages than physicians (as shown in Table 1), and usually invited physicians to the thread within 10 turns (see Figure 2). After inviting a physician into the thread, patients conversed with physicians for approximately 14 turns, with a long tail distribution (see Figure 3). In total, almost half of all threads had 21+ more turns across system, patient, and physician interactions (see Figure 4). The median time-to-resolution (calculated as the elapsed time from the first patient message in the thread to the first physician outcome note generation) is 12 hours.

Table 1: Thread Dataset Summary Statistics

Statistic	Value
# Patient Msgs	1,940
# Physician Msgs	1,339
# System Msgs	1,006
# Mean Turns (Total)	23.81
# Mean Turns (Patient - System)	9.95
# Mean Turns (Patient - Physician)	13.86

Additionally, we used OpenAI's gpt-4.1 with a few-shot example prompt to classify threads into mutually exclusive thread categories (i.e., Acute Care, Chronic Care, Lifestyle/Prevention, General Advice/Informational, Behavioral Health, Data Review, or Other). The majority of threads ($> 60\%$) were Acute Care or Chronic Care focused - more details about the thread category descriptions and statistics are made available in Appendix A.

2.1. Thread De-identification Process

To meet the "Safe Harbor" determination under Section 164.514(a) of the HIPAA Privacy Rule, we built a multi-stage pipeline to (1) replace PHI with realistic "surrogate" pseudonyms, and then (2) additionally paraphrase patient, system, and physician messages using a large language model. A full diagram of the workflow is shown in Figure 1.

PHI Surrogacy. We first used Microsoft Azure's Health De-identification Service to replace PHI throughout message threads with realistic "surrogate" pseudonyms (e.g., names of people, organizations, hospitals). Surrogate pseudonyms are kept consistent within message threads and thread outcome notes.

Thread Paraphrasing. To further improve patient anonymity, we used OpenAI's gpt-4.1 to paraphrase existing messages sent within a thread. In doing so, we sought to preserve the contents and messaging style of the original message, while ensuring patient/physician anonymity. The prompt used to perform the message paraphrasing is provided in Appendix B.

Validation and Post-Processing. After dataset creation, clinicians manually spot-checked threads to ensure that PHI elements were removed, and that the paraphrased threads still conveyed a similarly consistent clinical narrative. Patient images and uploaded multi-media files could not be de-identified, and are excluded from the dataset release.

2.2. Thread Outcome Notes Processing

For each thread in the dataset, physicians assigned to the thread were asked to draft up a "thread outcome" note prior to closing or resolving the thread. The thread outcome notes predominantly take the form of a Subjective, Objective, Assessment and Plan (SOAP) note, and are a standardized form of clinical documentation used to structure patient clinical encounters (Podder et al., 2025). Because of the long,

ongoing nature of threads, a thread can have multiple SOAP notes, with subsequent SOAP notes providing additional context into new clinical encounters (e.g., updates to lab results, updated findings, trajectory of symptoms). Given our asynchronous platform set-up allows for multiple chief complaints within a thread, some threads contain SOAP notes documenting different chief complaints. We provide the ordinal rank of where these SOAP notes occurred relative to the thread messages, such that dataset users can identify when temporally the SOAP notes were written relative to the ongoing thread messages.

Thread outcome notes were de-identified and paraphrased using a similar process to the thread messages. As shown in Figure 1, the thread outcome notes were first surrogated using the same Azure de-identification endpoint, and then paraphrased using a thread paraphrasing prompt (see Appendix B). Of note - surrogated PHI elements are consistent across thread outcome notes and their attached message threads (e.g., a patient pseudonym is the same across all thread outcome notes and within all messages in the message thread), making it easier for researchers who might want to reference surrogated names, locations, or other data elements across thread messages and notes.

2.3. Emergency Escalation Risk Label

Similar to HealthBench, we include a consensus emergency escalation risk label agreed-upon by at least two physicians (Arora et al., 2025). Physicians were asked to independently read through message threads, and label for whether the thread contained content (e.g., thread messages, medical profile and health record details, uploaded images and files) that warranted immediate, in-person care at an emergency department. Two independent physicians reviewed the thread language, along with contents from a patient's medical profile (which was not released) to assign escalation labels. Disagreements were adjudicated and resolved through consensus, ensuring that only threads with clear clinical risk were marked as escalatory.

3. Example Use Cases

We anticipate *HealthConvos* serving as a public benchmark dataset that can support a wide variety of natural language or human-AI interaction studies. In particular, we highlight a few targeted case-studies for evaluating agentic history-taking or assessing clinical note generation.

3.1. Evaluating agentic history-taking

A key motivation for *HealthConvos* is to enable rigorous evaluation of conversational agents designed for pre-visit intake or symptom triage. By aligning chatbot-prompted questions with the physician-initiated questions in the dataset, researchers can quantify whether an automated system asks clinically relevant follow-up questions at the right stage of the dialogue. In particular, these evaluations would allow structured comparisons of question ordering, coverage of differential diagnoses, and conversational efficiency (e.g., benchmarking LLM responses against real doctor's questions). Similar to the gatekeeper setup proposed by Microsoft, researchers could design a similar setup using this dataset to benchmark the performance of a multi-turn agentic history-taking agent (Nori et al., 2025).

3.2. Assessing clinical note generation

Because threads are paired with physician-authored SOAP notes, the dataset enables benchmarking of automatic clinical documentation systems. Researchers can investigate how well models can produce notes incrementally during a conversation, reflecting partial patient histories (given that many threads have multiple SOAP notes, HealthConvos provides information to identify when SOAP notes were created relative to the ongoing thread messages).

3.3. Identifying escalatory risk

Each conversation includes physician-adjudicated assessments of whether the case warranted escalation to urgent or in-person care. These annotations provide a valuable testbed for risk prediction models. Systems can be evaluated on their ability to flag concerning conversational signals early, supporting safer triage decisions in real-world deployments.

4. Discussions and Future Work

Unlike prior corpora that are largely synthetic or narrowly scoped, HealthConvos provides de-identified, multi-turn interactions derived from genuine patient-initiated threads. By pairing these threads with physician outcome notes and escalation risk annotations, HealthConvos supports the assessment of

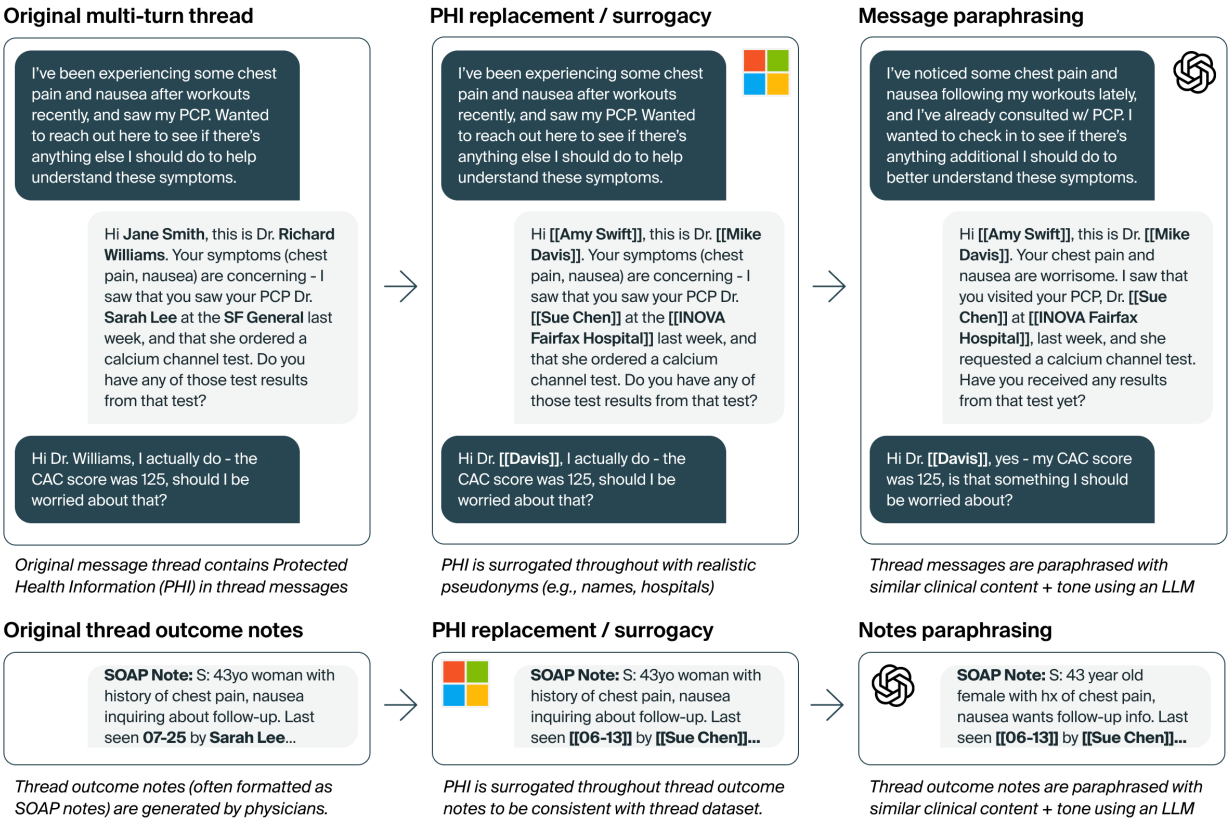

Figure 1: An example of the de-identification workflow showing (on the upper left) the original message threads, (in the upper middle) the process of replacing PHI elements with realistic surrogates, and (on the upper right) the paraphrasing of messages to maintain clinical content and tone. The same process was done for the thread outcome SOAP notes; surrogated PHI elements are consistent through thread messages and outcome notes.

history-taking agents, SOAP note generation, and emergency escalation risk.

Limitations. HealthConvos is limited to text-only messages and does not include the full depth of patients' medical records, as including such information would increase the risk of re-identification. Some physician diagnoses and SOAP notes were based on uploaded multimedia files (e.g., photos of rashes, lab results), which are also excluded from HealthConvos for the same privacy reasons. Consequently, agent and physician responses may rely on health record information that are not available in the dataset. Moreover, the platform's asynchronous communication model limits immediate comparisons to synchronous or in-person care settings; for example, patients on the Counsel platform often spend significant time composing longer messages before initiating a consultation with a physician, whereas in-person settings typically require patients to recall and convey symptoms more spontaneously.

Future Work. Based on community input and feedback on the HealthConvos GitHub repository, we aim to release additional threads and support additional use-cases, as de-identification and privacy allow. We also plan to explore integrating multimodal data and richer patient context to enable more realistic and clinically useful conversation scenarios.

References

A. Abbas, M. S. Rehman, and S. S. Rehman. Comparing the performance of popular large language models on the national board of medical examiners sample questions. *Cureus*, 16:e55991, 2024. doi: 10.7759/cureus.55991. URL https://doi.org/10.7759/cureus.55991.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025. doi: 10.48550/arXiv.2505.08775. URL https://doi.org/10.48550/arXiv.2505.08775.

Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H. Kann, Fallon E. Chipidza, Jonathan Leeman, Hugo J. W. L. Aerts, Timothy Miller, Guergana K. Savova, Jack Gallifant, Leo A. Celi, Raymond H. Mak, Maryam Lustberg, Majid Afshar, and Danielle S. Bitterman. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health*, 6(6):e379–e381, 2024. doi: 10.1016/S2589-7500(24)00060-8. URL https://doi.org/10.1016/S2589-7500(24)00060-8.

L. F. de Paiva, G. Luijten, B. Puladi, and J. Egger. How does deepseek-r1 perform on usmle? *bioRxiv*, 2025. doi: 10.1101/2025.02.06.25321749. URL https://doi.org/10.1101/2025.02.06.25321749.

Lars Masanneck, Linea Schmidt, Antonia Seifert, Tristan Kölsche, Niklas Huntemann, Robin Jansen, Mohammed Mehsin, Michael Bernhard, Sven G. Meuth, Lennert Böhm, and Marc Pawlitzki. Triage performance across large language models, chatgpt, and untrained doctors in emergency medicine: Comparative study. *Journal of Medical Internet Research*, 26:e53297, 2024. doi: 10.2196/53297. URL https://doi.org/10.2196/53297.

K. McCaffery, E. Cvejic, and J. Ayre. Use of ChatGPT to obtain health information in australia, 2024: insights from a nationally representative survey. *Medical Journal of Australia*, 222, 2024.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P. Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025. doi: 10.48550/arXiv.2506.22405. URL https://doi.org/10.48550/arXiv.2506.22405.

Vikas Podder, Victor Lew, and Salam Ghassemzadeh. Soap notes. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2025.

Marley Presiado, Alex Montero, Lunna Lopes, and Liz Hamel. Kaiser Family Foundation health misinformation tracking poll: Artificial intelligence and health information. Technical report, Kaiser Family Foundation, 2024.

K. Saab et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024. doi: 10.48550/arXiv.2404.18416. URL https://doi.org/10.48550/arXiv.2404.18416.

S. Wang, M. Hu, Q. Li, M. Safari, and X. Yang. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224*, 2025. doi: 10.48550/arXiv.2508.08224. URL https://doi.org/10.48550/arXiv.2508.08224.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. Mentalchat16k: A benchmark dataset for conversational mental health assistance. *arXiv preprint arXiv:2503.13509*, 2025. doi: 10.48550/arXiv.2503.13509. URL https://doi.org/10.48550/arXiv.2503.13509.

Appendix A. Thread Categorization

We categorized threads into mutually exclusive thread categories using OpenAI's gpt-4.1. The entire thread message history was used to classify the thread. The categories, their frequency, and their descriptions are as follows (the attached few-shot examples were used in the prompt to help improve classification accuracy):

Category	%
Chronic Care	36.9
Acute Care	29.5
Other	18.8
Behavioral Health	4.7
Data Review	4.7
Lifestyle/Prevention	2.7
General Advice/Info	2.7

General Advice / Informational Basic health questions driven by curiosity or a desire to learn, not tied to an existing medical issue. *Examples:* "Is it safe to take Tylenol after a night where I drank alcohol?"; "Can certain sports like swimming cause carpal tunnel to return?"

Acute Care Evaluation and management of a new or worsening discrete health issue. This category includes questions about whether a patient should seek urgent or emergent care, as well as "acute on chronic" conditions (e.g., a COPD exacerbation). *Examples:* sore throat, back pain, COPD exacerbation, migraines, chest pain.

Chronic Care Questions about an existing health issue. Includes requests for second opinions, referrals to specialists, or lifestyle advice (diet, exercise) that pertain to a specific chronic condition. Excludes behavioral/mental health issues and questions about medical data review (labs, imaging). *Examples:* managing high blood pressure, longstanding skin mole, chronic shoulder pain, best diet for COPD.

Lifestyle / Prevention Questions related to diet, exercise, sleep quality or habits, vaccines, or screening tests.

Behavioral Health Questions related to mood, anxiety, ADHD, therapy or psychiatric medications, or coping strategies.

Data Review Requests to review lab reports, imaging reports, or other objective medical record data.

Other Administrative questions, technical issues, empty or one-word threads without sufficient context, or messages that do not fit into any of the above categories.

Appendix B. Thread Message and Outcome Note Paraphrasing

Thread messages were paraphrased and rewritten using OpenAI's gpt-4.1. The prompt for rewriting thread messages was:

```
You are given the following message from a
↪ patient and are tasked with rewriting the
↪ patient's message in your own words,
↪ while keeping the contents, style, and
↪ tone of the original patient message
↪ similar. Do not return any text besides
↪ the paraphrased message.\n<Patient
↪ Message>{patientMessage}</Patient
↪ Message>
```

Thread outcome notes were paraphrased and rewritten using OpenAI's gpt-4.1. The prompt for rewriting thread messages was:

```
You are given the following physician (SOAP)
↪ note and are tasked with rewriting the
↪ note in your own words, while keeping the
↪ contents, style, and tone of the original
↪ note. Make sure to match the original
↪ formatting style of the note. Do not
↪ return any text besides the paraphrased
↪ message.\n<Physician
↪ Note>{physicianNote}</Physician Note>
```

Appendix C. Message Counts and Turns Per Thread

We visualize the number of turns attributable to patient-system interactions (corresponding to patients conversing with the system agent for history-taking) in Figure 2, and the number of turns attributable to patient-physician interactions (corresponding to patients conversing with the assigned

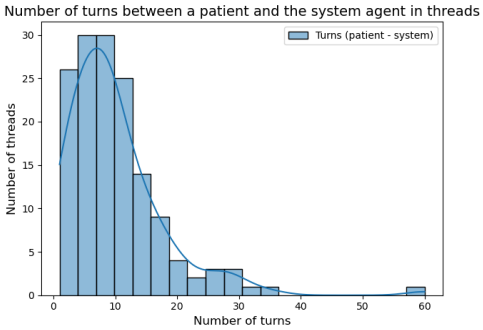

Figure 2: Visualizing the number of turns within each thread where the patient converses with the system agent.

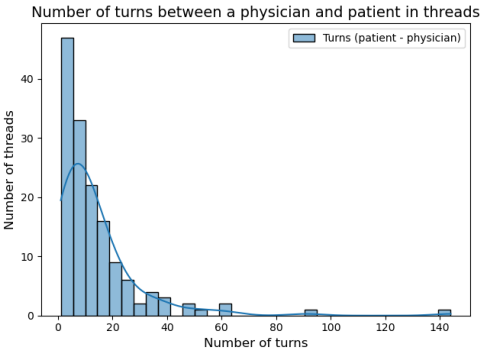

Figure 3: Visualizing the number of turns within each thread, where the patient converses with the physician.

physician) in Figure 3. We also show the total number of turns (measured as changes in message sender for blocks of messages, e.g., from system to patient, or from patient to physician) in conversations in Figure 4, as well as the number of discrete messages generated across threads in Figure 5.

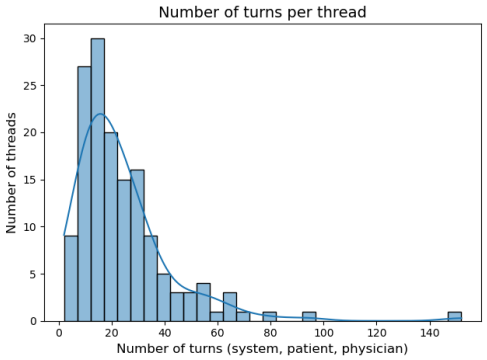

Figure 4: Visualizing the number of turns within each thread, with a turn representing a change in message sender (e.g., from system to patient, or from patient to physician).

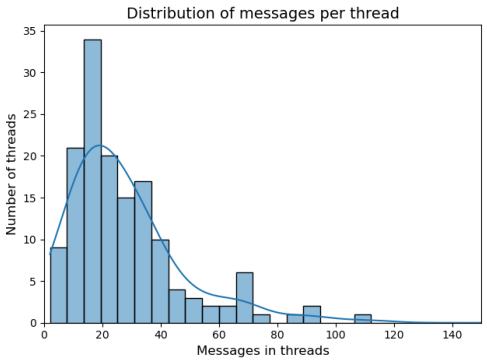

Figure 5: Many (43%) of threads were closed or resolved in under 20 messages, but some patients with more chronic issues engaged in longer conversations.