Title:

Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: how does it compare to traditional methods?

Short title: LLMs for thematic analysis

Authors:
Walter S Mathis, MD[1]
Sophia Zhao, BS BA[1]
Nicholas Pratt, MD RN[1]
Jeremy Weleff, DO[1]
Stefano De Paoli, PhD[2]

Affiliations:

1. Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut, USA
2. Division of Sociology, School of Business, Law and Social Sciences, Abertay University, Dundee, Scotland

Corresponding Author:
Walter Mathis, MD
34 Park St. Rm 162A
New Haven, CT 06519
T: (203) 314-9559
Walter.Mathis@Yale.edu

Declaration of interest: none

Word Count: 5,915
Tables/Figures: 2 tables, 4 figures
References: 48

Abstract

Large language models (LLMs) are generative artificial intelligence that have ignited much interest and discussion about their utility in clinical and research settings. Despite this interest there is sparse analysis of their use in qualitative thematic analysis comparing their current ability to that of human coding and analysis. In addition, there has been no published analysis of their use in real-world, protected health information. Here we fill that gap in the literature by comparing an LLM to standard human thematic analysis in real-world, semi-structured interviews of both patients and clinicians within a psychiatric setting. Using a 70 billion parameter open-source LLM running on local hardware and advanced prompt engineering techniques, we produced themes that summarized a full corpus of interviews in minutes. Subsequently we used three different evaluation methods for quantifying similarity between themes produced by the LLM and those produced by humans. These revealed similarities ranging from moderate to substantial (Jaccard similarity coefficients 0.44-0.69), which are promising preliminary results. We discuss their potential utility in qualitative research methods as refinement of these technologies continues.

Keywords: artificial intelligence, large language models, qualitative methods, thematic analysis, mental health

1. Introduction

Large Language Models (LLMs), a subset of generative Artificial Intelligence (AI), have been garnering increasing interest for their potential applications in both private and public sectors, sparking numerous discussions about their role in clinical and research environments (De Paoli, 2023) [1]. Although natural language processing (NLP) is already a recognized tool for analyzing clinical and research data across various domains [2-6], the growing body of research on LLMs is opening new avenues of exploration and raising questions about their potential benefits and uses.

LLMs have offered significant advantages for medical professionals in various domains ranging from efficiency, note-taking/documentation, research, medical education, and patient care [7-9]. They also show promise for addressing health systems weaknesses and improving health equity. For instance, LLMs can improve real-time language translation, facilitating communication between patients and providers who speak different languages [10]. LLMs have shown to be able to assist in answering questions that patients have, including generating health education materials or answers that are accurate, empathetic, and understandable [11-13].

In the field of medical research, there has been a substantial increase in publications related to AI's use in research methodologies and administrative tasks, covering nearly all medical specialties [14]. Thematic analysis (TA) is an important research technique in psychiatry and many other fields, primarily due to its proficiency in tackling types of questions that other methods struggle to address, especially in open-ended or exploratory research contexts. A key use of TA involves the qualitative examination of interview transcripts, aiming to uncover and interpret the underlying themes, thereby providing more profound insights [15]. The integration of LLMs into TA is appealing, particularly in the medical research context, due to the technical complexities and substantial resource demands, especially in human resources and time, associated with such analyses. The literature on leveraging LLMs as textual analytic tools in TA is growing [16,17]. Notably, one study has used the commercial LLM GPT-4 to undertake inductive thematic analysis of an existing interview corpus using Braun and Clarke's framework for TA [1].

While there are numerous approaches or frameworks for formally applying TA to a text corpus, the six-step approach outlined by Braun and Clarke is widely recognized and commonly used [15]. The six progressive phases start with familiarization with the data, then generating codes from the data, searching for themes, reviewing the themes, defining and naming the themes, and finally producing the report of the themes derived from the data. This approach is based on inductive coding where codes and themes are derived directly from the data itself. While designed for use by human researchers, certain phases of the process seem well suited for offloading to NLP or LLM tools. The emergence and progression of AI and LLMs have the potential to level the playing field for this method of research by substantially decreasing human resource requirements. This can be achieved by enhancing the proficiency of current experts and research teams, thereby enabling them to handle larger-scale projects, or by facilitating research endeavors that might be impractical for teams without the required infrastructure.

Despite the growing use of LLMs and the interest in further developing their use in qualitative research settings, a major limitation to incorporating them into research involving protected health information (PHI) is the dependence on cloud-based commercial services. Sending PHI offsite poses significant privacy, ethical, and security risks as well as statutory administrative complications with Health Insurance Portability and Accountability Act (HIPAA) (in the US), and other relevant legislation in other countries such as General Data Protection Regulation (GDPR) in the European Union. At this time, OpenAI is not signing Business Associate Agreements with HIPAA-regulated entities, essentially precluding applications involving any PHI [18]. In addition, the most capable LLMs at this time – GPT-4, Claude, Bard, Gemini – are all commercial and closed, where inference can only be performed by using their internet-based service.

Fortunately, the algorithm and data structure for training and then making inferences (getting output) from a model at the heart of nearly all LLMs and other recent generative AI innovations (including ChatGPT) is open source and free to use. This technology, called the transformer, was developed by Google Research in 2017 [19]. Additionally, the open-source community, including some very large businesses like Meta (formerly Facebook) have contributed massive amounts of effort and money to train models that are completely open source and free to use on local hardware [20]. Also, the sizes of these open-source models are generally engineered to be accommodated more readily on commercial hardware that an individual or lab might have.

The primary aim of the paper is to outline a workflow for producing codes and themes outlined in the Braun and Clarke framework using a locally-hosted LLM and advanced prompting strategies. To our knowledge, there are no studies examining the use of LLMs to augment the TA of real-world clinical health interviews containing PHI nor using local, open-source LLMs to do so, a gap in the literature we aim to close. We will use an interview corpus that has already been analyzed using conventional TA techniques and hence we will already have a control set of themes for comparison. By running the LLMs locally, we avoid the complications / limitations imposed by HIPAA onto PHI. The secondary aim is to explore techniques for assessing the performance of LLM-generated content in comparison to human-generated output.

2. Materials and Methods

For this study we will use prompt engineering and open-source on-site tools to produce codes and then themes from an existing corpus of semi-structured interviews with patients and clinicians. We will then compare these codes and themes with those produced from the same corpus via more traditional human-based TA methods. We will use three different approaches for comparing the products of the different methods. This study was deemed exempt from IRB review by the Yale University Human Research Protection Program.

Interviews

The primary data source for this study is a set of 21 qualitative interviews, 14 with clients and 7 with clinicians treating those clients, conducted at a community mental health center in Connecticut, as previously described [21]. This primary study was an exploration of the barriers and facilitators of primary care access among an Assertive Community Treatment (ACT) team. The clients were patients with severe mental illness and the clinicians were social workers, therapists, and nurses with bachelor's or master's degrees. They were recruited based on availability. Eligible participants were sufficiently fluent in English to understand the consent form and participate in an interview.

The semi-structured interviews included several predetermined questions to explore themes related to barriers to primary care (see supplement). Follow-up questions were asked to probe deeper into issues brought forward by the participants. These interviews were recorded on a handheld digital recording device. Client interviews had a median duration of 10.5 minutes and once transcribed had a median word count of 1560. Clinician interviews had a median duration

of 19.2 minutes and median word count of 3244. The relatively short duration of client interviews reflected their variable verbosity during of comfort level with the interview setting.
In addition to the interviews themselves, conventional TA had already been performed by a human researcher as part of the primary study for which these interviews were collected [15]. For the current study, these human-produced themes would serve as a basis for comparison for those produced by our LLM approach. In both human- and LLM-produced approaches, client and clinician interviews were analyzed separately but using identical methods.

Tools

Whisper

To convert the recorded audio of the interviews into digital text, we used an open-source automatic speech recognition model called Whisper [22]. It translates the complex patterns in audio waveforms into textual data by leveraging models trained on vast datasets. This training enables it to understand various accents, languages, and even noisy environments. Of note, to simplify our data flow, we converted the text of the entire interview – both interviewer and interviewee – into a single file with no speaker diarization (delineating who was saying what).

Llama 2

In this study, we utilized the LLaMA-2-70B-Instruct [23] as our LLM. This is a specialized version of the Llama 2 70 billion parameter foundation model – trained and released free to the public by Meta (formerly Facebook) (https://ai.meta.com/llama/) – that has been fine-tuned on instruction-based prompts for more precise control and higher accuracy in generating text-based responses. Llama 2 represents a significant evolution in natural language processing (NLP), offering enhanced understanding and generation capabilities. Its large-scale parameter architecture contributes to its ability to grasp complex concepts, discern nuanced contexts, and produce highly relevant and coherent text outputs.

We deployed the Llama 2 model on a local Ubuntu Linux server equipped with dual Nvidia RTX 3090 GPUs with a combined VRAM of 48GB, which provided the necessary computational power for handling the model's extensive requirements. We utilized a 4-bit quantized GPTQ version of the model, which allowed for optimized model size without significant loss in performance. The model was loaded using the ExLlama library, a specialized tool designed for efficient handling and operation of quantized Llama models [24] and we interacted with the model from Python scripts using an API [25] which facilitated batch processing of interviews and easy storing of outputs. We used hyperparameters (model settings) optimized for this task – balancing consistency and contextual insight (temperature: 1.31, top_p: 0.14, repetition_penalty: 1.17, top_k: 49) [26][1]

[1] 'Temperature' controls the randomness in the model's predictions, with higher values leading to more varied and creative outputs, and lower values resulting in more predictable and conservative responses. 'Top_p' specifies a threshold to sample from the most likely words; the model only considers the smallest set of words whose cumulative probability exceeds this threshold. 'Repetition penalty' discourages the model from repeating the same words or phrases. 'Top_k' limits the number of highest-probability vocabulary tokens to be considered for each step of the generation.

Sentence-T5-xxl

One approach we utilized for evaluating the similarity of LLM-produced and human-produced themes was Sentence-T5-xxl. This language model is specifically designed to handle a variety of sentence-level natural language processing tasks and has been optimized for capturing complex semantic information from text. It is the most robust model of this type currently available, having been trained on 2 billion sentence pairs and having 11 billion parameters [27]. To quantify the similarity between two sentences, the model converts each sentence into vector spaces – a mathematical representation of the sentence encoded with the semantic understanding the model has acquired during its training phase. Once, in their respective vector spaces, the similarity of the two sentences can be calculated using a cosine similarity. The result is a number between 0 and 1, with 1 being the maximum similarity. This approach is very fast, symmetrical (same score if comparing A vs B as B vs A), and highly reproducible (same score every time). The output is continuous (not binary), so additional analytics are required to determine an appropriate threshold value (> certain score means sufficiently similar).

Human analysis / theme generation

Prior to the current study, the same 14 client and 7 clinician interviews were transcribed and analyzed for themes by a human researcher [21]. De-identified transcripts were analyzed using an open coding procedure where text reflecting barriers to primary care were organized into discrete codes. Then, relationships between these codes were identified and they were categorized under overarching themes (e.g., “Attitudes,” “Motivations,” “Logistics and Resources”). Codes and themes were iteratively refined throughout the interview process. After each revision, previously coded interviews were reviewed and adjusted as needed. Initial codes discovered during the analysis of client interviews were utilized as a starting point in the analysis of clinician interviews; the same process of iterative review and refinement was repeated. Where needed, the interview guide was amended to address emerging themes.

LLM analysis / theme generation

In this preliminary study, we will not cover the entire six-step thematic analysis process outlined by Braun and Clarke. Similar to previous work [1], our attention will be focused on the Phases 2) the generation of initial codes, 3) the identification of themes, and 5) their subsequent definition and naming. We have chosen to focus on these steps because they best match the text summary and abstraction strengths of LLMs and would be particularly beneficial in augmenting the efforts of human thematic analysts, offering a complementary blend of machine efficiency and human insight.

The final phase of Braun and Clarke’s framework entails the writing up of the results, which we are not performing here with the LLM. Although De Paoli (2023) has shown that it is possible to use the results of a TA produced by an LLM also to write-up intermediate textual outputs, such as user personas [1]. However, for this manuscript it is not relevant to perform the write up phase.

The output produced by LLMs can significantly differ in terms of content, quality, and veracity, depending on the formulation of the task request, namely the construction of the prompt. A prompt is the set of instructions we request the LLM to perform.

For example, here is a prompt and response from ChatGPT:

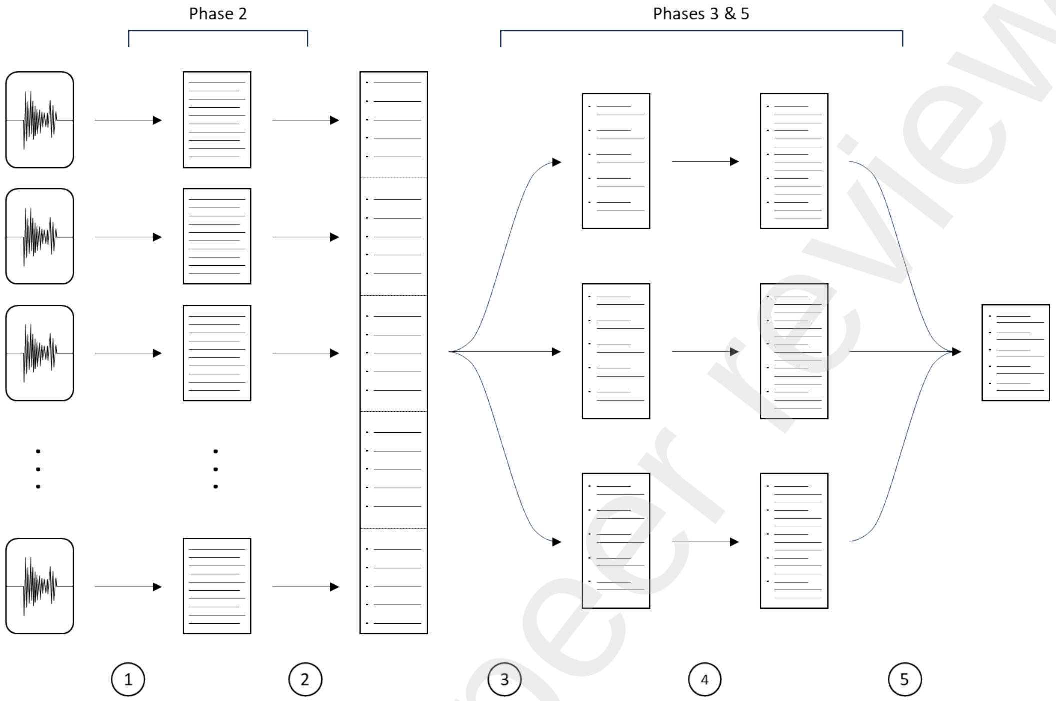
> **You:**
> Write the names of the three greatest American writers.
>
> **ChatGPT:**
> The three writers often regarded as among the greatest in American literature are:
>
> 1. Mark Twain
> 2. Ernest Hemingway
> 3. F. Scott Fitzgerald

A well-crafted prompt can lead to more accurate, relevant, and coherent outputs, effectively steering the model to fulfill the intended task. Conversely, a vague or poorly structured prompt can result in ambiguous or off-target responses. The crafting of a prompt which produces the desired responses from an LLM is often defined as "prompt engineering" [28]. Effective "prompt engineering" involves the strategic use of keywords, context setting, and clear instructions to activate the model's knowledge in a specific domain, thereby enhancing the utility and precision of the generated content. These techniques cannot be derived from the LLM itself but instead must be discovered empirically, though sometimes with the help of LLMs [29]. Hence the state of the art is always progressing. At this moment, one of the approaches that appears most promising is a prompt engineering technique called SmartGPT which combines several prompting techniques into a multi-step algorithm:

> *Iterative refinement.* Because output from LLMs is not deterministic, they can produce different answers with the same prompt (and we have no way of knowing which is best). This approach involves collecting multiple iterations of answers to the same prompt for further comparison [30].
>
> *Chain of thought.* This prompting technique involves structuring a prompt to encourage the LLM to articulate its reasoning process step-by-step, much like a human would when solving a problem out loud. This approach is particularly useful for complex reasoning tasks as it makes the model's thought process transparent, potentially improving the quality and accuracy of the answers [31].
>
> *Reflexion*. This prompting technique is designed to elicit linguistic feedback by having the LLM engage in self-reflection on its own output. Specifically, this method entails instructing the LLM to assess its own answer for any potential flaws [32].
>
> *Dialog-Enabled Resolving Agents.* A technique that improves the quality of the answers by incorporating the feedback from the Reflexion step and choosing the best output options [33].

The following outlines our approach for implementing the thematic analysis tasks corresponding to Clarke and Braum phases 2,3, and 5. See Figure 1 for visualization and prompt texts used.

Fig 1. Flowchart for theme generation. Step 2 corresponds to Braun and Clarke's Phase 2, generating initial codes. Steps 3,4, and 5 combine to deduce the highest quality themes from these codes, corresponding to Braun and Clarke's Phase 3 and 5, searching for themes and defining and naming themes.



Step 1. Whisper converts MP3 recordings to transcripts.

Step 2. LLM generates codes from transcripts, concatenates all codes.
Prompt: "Identify all themes in the text, provide a name for each theme in no more than 5 words, a condensed description of the theme, and a quote from the interview that supports the theme."

Step 3. LLM generates three iterations of themes from codes.
Prompt: "Determine how all the topics in the list of topics can be grouped together. Topics can be in more than one group. Provide a name and description for each group."

Step 4. LLM evaluates themes for flaws.
Prompt: "List the flaws and faulty logic of each answer option.
Let's work this out in a step by step way to be sure we have all the errors:"

Step 5. LLM resolves flaws and generates best list of themes.
Prompt: "You are a resolver tasked with finding the answers that best determines how all the topics in the list of topics can be grouped together.
1) removing any redundant or duplicate answers.
2) improving the answers based on the analysis of flaws
3) printing the improved answer in full
Let's work this one out in a step by step way:"

Phase 2: generating initial codes:

For each interview, the LLM was asked to identify all codes within the interview, provide a name, description, and quote to support it from the interview text. Because the term “code” has special meaning within TA, we used the term “theme” when prompting the LLM in this phase.

> **Prompt:**
> Identify all themes in the text, provide a name for each theme in no more than 5 words, a condensed description of the theme, and a quote from the interview that supports the theme.

All codes for client and clinician interviews were concatenated into a master list of codes for each. These lists were not de-duplicated, meaning even duplicate or very similar codes were included. For the python code used in this step, please see the Supplement.

Phase 3 and 5: Searching for, defining, and naming themes:

All client codes and clinician codes were then passed, separately, to the LLM via a series of prompts that ask to make themes from them. As this is a more abstract and sophisticated task than the previous one, we leveraged the SmartGPT prompt engineering approach described above to maximize the quality of the results. First the model was asked to identify themes from the lists of codes. As ‘themes’ had been used in the previous step to identify thematic analysis ‘codes’, we instead used the term ‘topic’ in the prompt.

> **Prompt:**
> Determine how all the topics in the list of topics can be grouped together. Topics can be in more than one group. Provide a name and description for each group.

As outlined in the SmartGPT approach, this step was iterated three times. Next, the LLM was asked to examine the three outputs for flaws.

> **Prompt:**
> List the flaws and faulty logic of each answer option.
> Let’s work this out in a step by step way to be sure we have all the errors:

Finally, the LLM was given all the proposed themes and all the assessments of their flaws and asked to address the flaws and generate the answers that best address the task.

> **Prompt:**
> You are a resolver tasked with finding the answers that best determines how all the topics in the list of topics can be grouped together
> 1) removing any redundant or duplicate answers
> 2) improving the answers based on the analysis of flaws
> 3) printing the improved answer in full.
> Let’s work this one out in a step by step way:

The final output from this last step is a list of themes, each with a name and a description.

Comparison of human and LLM theme generation

The assessment of similarity between human-generated and LLM-generated themes involves a two-step process. First, each theme produced by humans is compared with each theme generated by the LLM to determine thematic resemblance. Full theme name and description are used. This comparison results in a binary similarity matrix where each row is a human-produced theme, each column an LLM-produced theme, and each cell of the matrix has either a ‘1’ indicating similarity or a ‘0’ indicating a lack of similarity. Subsequently, this matrix is used to quantify the overall similarity between the two sets of themes.

We employed three distinct methods to produce the binary similarity matrices. These three approaches were applied to the client themes and the clinician themes separately.

Approach 1: Human Evaluation of Theme Similarity
This approach involved hand coding by three independent coders. Each coder compared every LLM-generated theme with each human-generated theme, determining whether there was significant similarity or overlap. The coders were blinded to the source of the themes. They each produced a binary similarity matrix and from these individual matrices, we created a consensus matrix – a theme pair was considered similar if at least two out of the three coders agreed on its similarity. Cronbach’s alpha was computed between raters.

Approach 2: Sentence-T5-xxl Evaluation
We utilized Sentence-T5-xxl, a machine learning model adept at generating semantically meaningful sentence embeddings. By measuring cosine similarity, we quantified the semantic closeness between every pairing of human-generated and LLM-generated themes. Since cosine similarity yields a continuous score (typically ranging between 0.6 and 0.9), we conducted a sensitivity analysis to establish a threshold for similarity. This threshold enabled us to translate the continuous scores into a binary system (similar or not similar).

Approach 3: LLM-Based Evaluation
In this method, we re-engaged the same LLM used for theme generation. We fed each pair of LLM-generated and human-generated themes back into the LLM with a specific prompt: “Do these themes overlap? Answer with ‘yes’ or ‘no’ only.” This allowed us to gauge the LLM’s perspective on the overlap between its own thematic output and that created by humans.

To quantify overlap of LLM-produced and human-produced themes, we computed the Jaccard similarity coefficient for each binary similarity matrix produced by each of the above methods. This coefficient is a measure of similarity between two sets and is defined as the size of the intersection divided by the size of the union of the two sets.

Finally, to quantify the overlap of the outputs of each method, we used the Jaccard similarity coefficient to compare the binary similarity matrices between them.

3. Results

Human analysis / theme generation

Thematic analysis conducted by the human researcher identified a total of 96 codes from the 14 client interviews and 180 codes from the 7 clinician interviews. After removing or merging duplicate or very similar codes, 31 unique client codes and 43 unique clinician codes remained.

Subsequent TA of these codes produced 6 overarching themes within the client data and 7 themes within the clinician data. The names and descriptions of these themes are presented in Table 1.

Table 1. Human-coded themes and LLM-coded themes for client and clinician interviews

Human Themes - Clients	**LLM Themes - Clients**
1) Logistics & Resources - Clients require transportation to get to and from appointments. - Client relies on communication technology to be notified of appointments. 2) Knowledge - Client obtains advice and guidance from sources outside of the primary care provider (e.g. from online, through family relationships). 3) Attitudes - Client prefers not to know about their health complications. - Client believes that medical care is a privilege. - Client values primary care based on upbringing. - Client believes they are at low risk of serious physical illness. - Client feels that engaging with primary care is an invasion of independence. - Client believes that primary care providers do not center patients in their care (e.g. providers prioritize money over patient well-being). 4) Motivations - Client gives importance to building good relationships and trust with their provider. - Client believes that the primary care provider is a good source of knowledge to stay healthy. - Client believes that primary care is a way to connect to specialists. - Clients are motivated to take care of their physical health (e.g. due to threat of potential disease, guilt, desire to make progress).	1. Healthcare Access and Support Systems: This group includes topics related to the challenges, barriers, and facilitators involved in accessing healthcare services such as appointment scheduling, transportation, and support systems like the ACT team. It also covers the impact of work on healthcare access and the help received from family members. 2. Interactions with Medical Professionals: This group focuses on the interactions and relationships between patients and medical professionals including trust, communication, satisfaction, and previous experiences. It also encompasses the importance of having a primary care provider and the quality of interaction with doctors. 3. Self-Care and Prevention: This group emphasizes the individual's role in taking charge of their health through self-care practices, prevention methods, screenings, exercise, diet, and education. It highlights the significance of prevention and the participant's motivation to improve their health. 4. Specific Health Conditions and Concerns: This group addresses particular health concerns such as chronic pain management, dental anxiety, x-ray aversions, and mental health issues. It also explores experiences with substance abuse and physical illness. 5. Overall Health and Wellness: This group revolves around general health maintenance, water intake, checkups, and the importance of primary care. It also touches upon racism experienced within the healthcare system.

5) Experiences - Client previously experienced discrimination or traumatic events in a medical setting. - Client does not feel heard by their provider. - Client is overwhelmed by medical care. - Client is overwhelmed by medical environments. - Providers do not simplify information enough to client. 6) Abilities - Client has competing priorities that prevent them from being able to prioritize primary care. - Co-morbid psychological symptoms prevent client from using primary care.	6. Physical Health: This group is focused on physical health, which includes experiences with physical illness, the role of exercise, and the significance of primary care in maintaining physical health.
Human Themes - Clinicians	**LLM Themes - Clinicians**
1) ACT Team Capacities - Faulty record-keeping prevents the ACT team from directing client toward primary care. - ACT team provides client assistance or support in navigating primary care. - ACT team (adequately or inadequately) normalizes primary care for client. - ACT team respects and builds rapport with client, influencing primary care decision making. 2) Logistics & Resources - Clients require transportation to get to and from appointments. - Insurance limitations prevent client from accessing certain physical health services. - Primary care availability may not fit client's own schedule or physical health needs. 3) Knowledge	1) Client Factors: This group includes factors related to individual clients, such as their experiences, behaviors, and perspectives. 2) Service Provision Factors: This group focuses on aspects of service provision, including institutional practices, staff training, and quality of care. 3) Structural Barriers: This category covers barriers to accessing care at an institutional level, such as financial constraints, insurance coverage, and documentation requirements. 4) Family Influences: This group highlights the role of families and significant others in shaping client behavior and decisions around healthcare.

- Client lacks understanding of the benefits of primary care. - Client lacks understanding of the equal importance of both physical and mental health. 4) Attitudes - Client is scared of medical intervention. - Client prefers not to know about their health complications. 5) Motivations - Client relies on external motivation (food, money) to use primary care. - Client lacks support from family members to use primary care. - Client lacks interest in and motivation to use primary care. - Aspects improve client motivation to use primary care (gaining confidence, getting comfortable, facing imminent threat of disease). 6) Experiences - Client previously experienced discrimination or traumatic events in a medical setting. - Providers misunderstand client because of their mental illness or substance abuse disorder. - Client is overwhelmed by medical care. - Positive experiences w/ PCPs. 7) Abilities - Client has competing priorities that prevent them from being able to prioritize primary care. - Client's life circumstances result in physical health disparities. - Co-morbid psychological symptoms prevent client from using primary care. - Psychotropic medications worsen client's physical health.	5) Promotional Approaches: This group emphasizes strategies aimed at increasing utilization of primary care services, like building rapport, expanding services, and educating clients.

These themes encapsulate a broad spectrum of factors influencing healthcare experiences and perceptions. They cover logistical and resource challenges in accessing care, the role of external knowledge sources, and diverse attitudes towards health and medical care, including avoidance, privilege perception, and skepticism about provider motives. Motivations for engaging with primary care range from relationship-building with providers to external incentives and health threats. Experiences highlight past negative interactions in medical settings, feelings of being overwhelmed, and communication barriers. Abilities are constrained by competing life priorities and psychological symptoms. Additionally, themes reflect the capacities of ACT teams in supporting clients, the impact of logistics and knowledge gaps on healthcare access, and the complex interplay of client attitudes, motivations, and abilities in navigating healthcare amidst life challenges and medication effects.

LLM analysis / theme generation

In step 2, the LLM produced a total of 88 codes from the 14 client interviews and 51 codes from the 7 clinician interviews. In steps 3-5 these codes were refined to 6 client and 5 clinician themes (Table 1).

These themes encompass a comprehensive look at healthcare experiences, covering the practicalities and challenges of accessing healthcare, the nature of patient-medical professional relationships, the role of self-care and prevention in health management, specific health concerns including chronic conditions and mental health, and the impact of broader factors such as family influence, institutional practices, and structural barriers on healthcare experiences. Additionally, they include perspectives on overall health and wellness, physical health maintenance, client-specific factors, service provision quality, and strategies to enhance primary care utilization.

Comparison of human and LLM theme generation

Human coding of theme similarity

Three human coders independently produced binary similarity matrices between human- and LLM-produced themes and two consensus binary similarity matrices were derived from these (Fig 3). The challenge of assessing thematic similarity is highlighted by the modest agreement between the three coders. Cronbach alpha scores revealed a moderate level of agreement among the three binary similarity matrices of the different human coders, with a scores of 0.72 for client themes and 0.61 for clinician themes. In many social science research contexts, a Cronbach's alpha of 0.7 is considered an acceptable level of internal consistency, while a score above 0.8 is deemed good, and above 0.9 is excellent.

For client themes, the Jaccard similarity coefficient derived from the consensus binary similarity matrix was 0.44, indicating a moderate degree of similarity between the human- and LLM-produced themes (Table 2). For clinician themes, the coefficient stood at 0.51, reflecting a slightly higher, yet still moderate, level of similarity.

Sentence-T5-xxl coding of theme similarity

After utilizing the Sentence-T5-xxl model to calculate cosine similarity scores for themes produced by humans and LLM, we conducted a sensitivity analysis to establish thresholds for converting these scores into a binary similarity matrix. The chosen thresholds were 0.75 for client themes and 0.725 for clinician themes, as detailed in the Supplement, and resulted in the binary similarity matrices depicted in Figure 4. The resulting Jaccard similarity coefficients were 0.56 for client themes and 0.69 for clinician themes, signifying different degrees of similarity (Table 2). The higher scores observed in clinician themes suggest a greater congruence in thematic content between human and LLM outputs in the clinician context.

LLM coding of theme similarity

Tasking the LLM with binary similarity assessment produced a third pair of binary similarity matrices (Figure 4). In assessing the extent of similarity between human- and LLM-generated themes using the language model, the Jaccard coefficient was found to be 0.58 for client themes and 0.63 for clinician themes (Table 2). This indicates a moderate degree of similarity for both sets, reflecting a significant overlap but also considerable difference.

Similarity between evaluation methods

Finally, we calculated Jaccard similarity coefficients between the binary similarity matrices generated by the three evaluation methods to assess their congruence. The results revealed a spectrum of similarity scores that defy easy generalization. For instance, the comparison between the human-graders method and Sentence-T5-xxl yielded a Jaccard coefficient of 0.64 for client themes, indicating a moderately high similarity, but only 0.45 for clinician themes, pointing to a moderate level of similarity (Table 2). In contrast, the alignment between human-graders and LLM methods was moderate for client themes, with a coefficient of 0.54, but was more similar for clinician themes with a coefficient of 0.67.

Fig 2. Results of human consensus coding for similarity of LLM-generated themes (bottom) to human-generate themes (left).

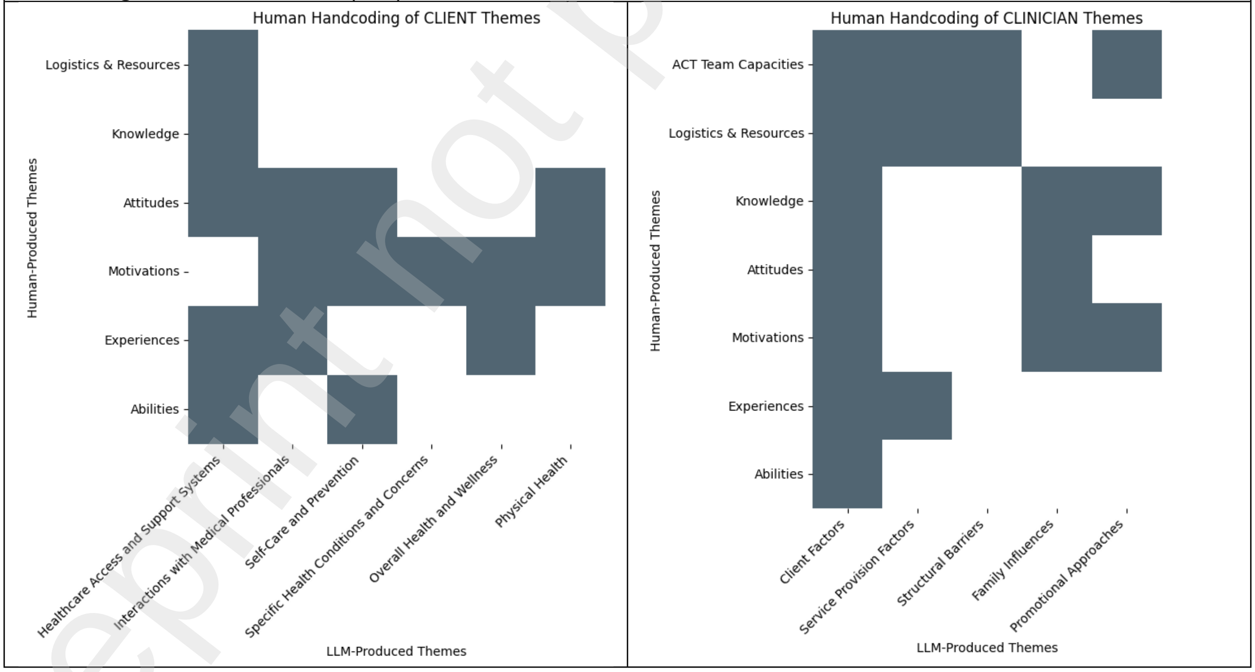


Fig 3. Results of sentence-t5-xxl cosine similarity coding of LLM-generated themes (bottom) to human-generated themes (left) overlaid with frames denoting which cells matched or exceeded similarity threshold of 0.75 for clients and 0.725 for clinicians.

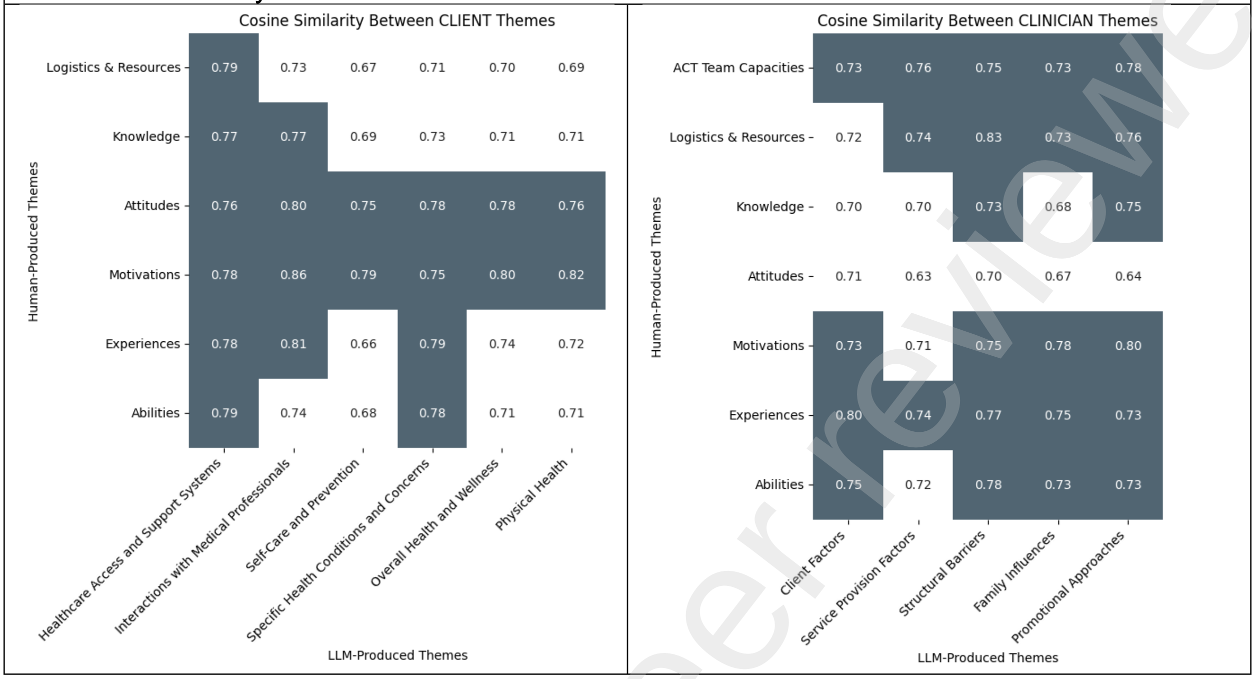


Fig 4. Results of LLM coding for similarity of LLM-generated themes (bottom) to human-generate themes (left).

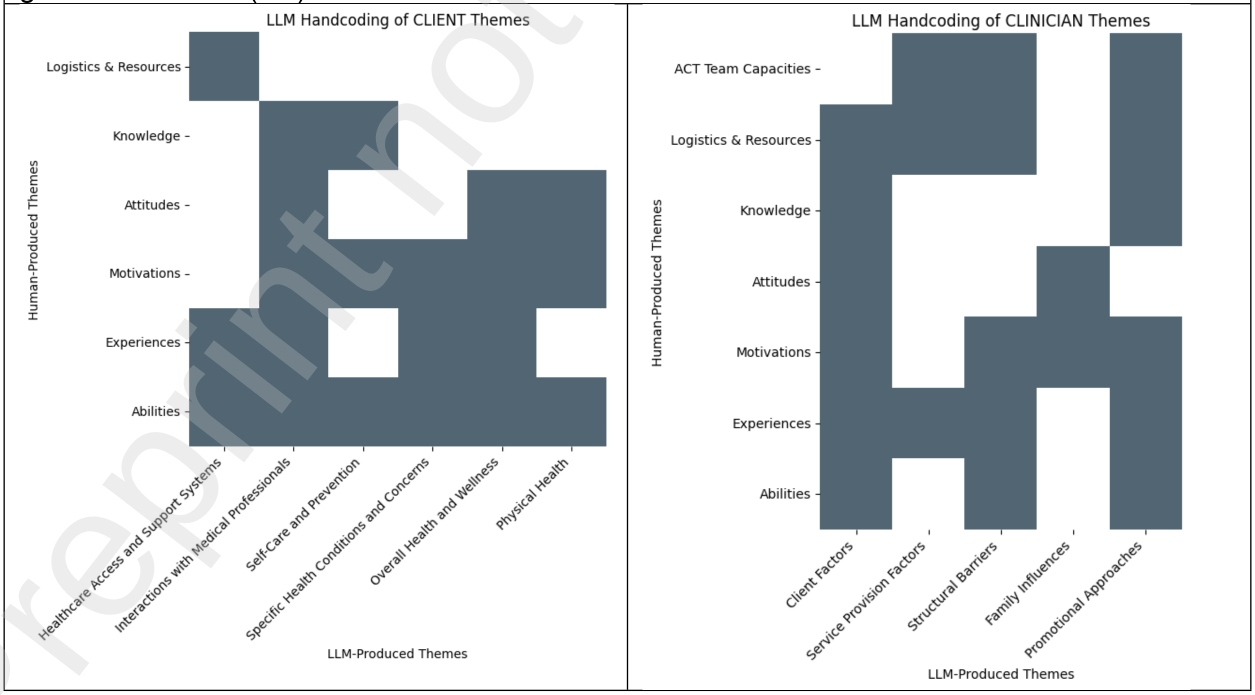


Table 2. Jaccard similarity coefficient scores for and between evaluation methods

		Self (Internal)	vs. hand	vs. sentence-t5-xxl	vs. LLM
Client	Hand	0.44	--	0.64	0.54
	Sentence-t5-xxl	0.56	--	--	0.58
	LLM	0.58	--	--	--
Clinician	Hand	0.51	--	0.45	0.67
	Sentence-t5-xxl	0.69	--	--	0.64
	LLM	0.63	--	--	--

4. Discussion

Here we present early findings that open-source LLMs can be used in the code and theme generation phases of TA for datasets containing PHI. Additionally, we have put forward three different evaluation methodologies for quantifying similarity between theme output between different methods.

Using these evaluation methodologies, we find that not only do LLM-based approaches produce robust themes that pass face validity but that quantitatively they overlap well with those themes produced completely and independently by traditional TA performed by a human. Using our most conservative evaluation method – hand coding by a team of human graders – yielded binary similarity matrices with Jaccard similarity coefficients of 0.44 for client themes and 0.51 for clinician themes, representing moderate to substantial level of similarity between the sets. We feel this is quite promising given the parallel methods used to produce the themes (human vs LLM) and the nearly infinite space of possible outputs. Also, it is worth noting that two (or more) human researchers performing inductive theme generation on a corpus of qualitative interviews are also very likely to produce different themes, each list equally valid. This is because inductive qualitative analysis is interpretative and depends to a greater degree on the analysts' sensibilities and unspoken assumptions [34].

We feel that this study is a major step in validating the idea that open-source LLMs are or could be powerful tools for TA, extending this class of tools even into projects where PHI is involved. Freeing the way for incorporation of these tools more broadly has the potential to massively impact TA research. TA is an incredibly resource-intensive methodology performed by a small group of highly specialized researchers. A recent study found that using NLP tools in the TA process saved 36 hours of research time and $1,100 for transcription costs [35]. Offloading some of the most resource-intensive phases of TA, the phases we covered in our current study, can greatly expand the reach of TA in several ways. First, research labs that are already performing TA can potentially extend their expertise and human resources to perform more studies than they previously could. Second, research groups who previously did not have the resources to bring to their study might now be able to proceed. Projects previously conceived of as being too modest to warrant TA might now be considered. Time sensitive projects might now fit into tight time windows with these approaches significantly speeding up the project.

In our examination of different evaluation techniques, we gained valuable insights across various methods. Hand-grading, the most resource-intensive approach, highlighted the inherent difficulty in achieving perfect reliability when comparing the similarity of theme sets. This was evidenced by our experiment where three human coders, unaware of the origin of the themes, compared each LLM-generated theme with human-generated ones. The reliability of their assessments, measured by Cronbach's alpha, was 0.608 for clinician themes and 0.720 for client themes. These scores underscore the challenge in this task, considering that in social science research, a Cronbach's alpha of 0.7 is usually seen as acceptable, above 0.8 as good, and above 0.9 as excellent.

The Sentence-t5-xxl method stood out for producing continuous (not binary), highly consistent, and reproducible data. It is very fast (less than a second) and allows for the fine-tuning of thresholds through sensitivity analysis.

In contrast, the LLM hand grading approach posed unique challenges, particularly in its suitability for grading its own output. LLMs struggle with quantifying themes numerically, although they manage binary assessments more effectively. This method took two minutes for a complete comparison and was notably sensitive to the phrasing of prompts (e.g., “Are these themes related” produced very different results from “Are these themes similar”). It has the advantage of not having to add anything to the tech stack (already up and running from theme generation).

Given the resource requirements of human coding, one might want to choose from the alternative approaches. When comparing these methods to human coding, the alignment varied based on the theme type. The LLM and human hand grading showed a good overlap for clinician themes but less so for client themes. Conversely, Sentence-t5-xxl and human hand grading aligned well for client themes, but the correlation was weaker for clinician themes. These findings suggest that the choice of the most suitable evaluation technique is contingent on its closeness to the accuracy and reliability of human coding.

As LLM technologies continue to improve – and open-source versions released to the public – the quality of the output will also improve. With this moving target, we feel that the evaluation and comparison of the methods we outline will continue to be important.

Limitations

There are several limitations to the current study. First and most fundamentally is the inherent difficulty of comparing the thematic output of two different approaches. While convention might make one more trusting of the themes produced by human researchers in the conventional method, one cannot treat them as exhaustive ground truth against which to compare the LLM-produced themes. Hence while we believe our similarity qualifications to be the best available evaluation methods, their interpretability has limitations. They suggest that these methods are converging on shared understandings of the interview corpus, but cannot say, for instance, that a theme produced by the LLM is “right” or “wrong”.

Of note, the primary data used in this study are from clients with serious mental illness who have lived lives of the margins of society with varying degrees of economic and educational advantage. These contexts may lead to interviews of shorter duration and less organization – perhaps complicating the task of thematic analysis – but also bring the invaluable perspective of their lived experience. We paired client interviews with those of clinicians who are mostly mental health professionals with advanced degrees, and we show similar results between them. But it should be noted that the generalizability of these findings should be considered when using these techniques in other settings. For the fullest understanding we would want to triangulate these data with a broader context – patient experiences, institutional history, cultural views of mental health, among others.

In addition, both human and LLMs bring in sources of bias in identification of codes and themes. There have been raised concerns about the propensity of LLMs to perpetuate biases in healthcare [36-38], with some studies showing ChatGPT models showing limited biases [39,40]. It is important to realize that at its foundation, clinical algorithms and prediction models have been criticized [41-43], and biases in NLP and machine learning (ML) approaches have been observed, particularly along racial lines, [44,45], but also gender/sex [46]. Of course, these limitations are no limited to LLM-based approaches as thematic analysis and interpretative research more generally inherently introduces bias [47].

transcend this study and apply to qualitative research broadly.

Another limitation is a technical one. This study was set up to use the premier open-source LLM at the time – a 70 billion parameter model. Even though we have hardware infrastructure beyond the norm (48GB of GPU RAM), we still had to run this model in a 4-bit quantized state (in essence each parameter rounded from 32-bit precision to 4-bit to fit inside the memory confines of our computer). While this is a common practice in limited resource settings, it is unclear exactly what impact on output quality quantization has on specific tasks – which seems to be minimal but not zero [48].

We must also acknowledge a limitation to the dissemination of the methods of this study is the considerable hardware, software, and technological acumen required to implement it. This is common to projects in the early stages of new technologies, but we think there are steps to address this. This overhead encompasses two key components: the infrastructure, which includes both the hardware and software required to operate the LLM, and the software designed for efficient interaction with the LLM. The infrastructure is, in essence, a one-time setup expense (money and time) that a lab or university could invest and then use across projects and share between workgroups. And, for many these startup pain points still be much more feasible that the human infrastructure required for traditional TA.

Secondly, the scripts used for this interaction could be integrated into a more user-friendly, drag and drop interface, moving away from the complexity of raw Python code. Once this integration is achieved, it would be straightforward to direct these user interfaces either to a locally hosted LLM or to a commercial, cloud-based LLM. This choice could be made simply with the click of a button, tailored to the specific requirements and data privacy concerns (like PHI) of the project.

Future work in this space would include these efforts to reduce the technical overhead as well as continued efforts to improve the quality of the LLM code and theme generation. One front for this would be monitoring as newer, better trained open-source models are released, which happen multiple times per year. Another would be staying up to date with prompt engineering techniques as they continue to improve the quality of the outputs from LLMs. Additionally, while we have limited the scope of this study to Phases 2, 3, and 5 of TA, in future work we would hope to expand into other phases such as the generation of initial TA reports. Such a report, produced on the fly for a clinician, has the potential to significantly impact real-time interventions informed from qualitative data.

5. Conclusion

Here we show fair concordance between human and LLM TAs within a corpus of semi-structured interviews. Though at early stages, this shows some promise but also some important limitations. We describe here how LLMs can be used in specific stages of TA, such as code and theme generation, which may have the potential to unload a significant amount of human labor in research processes and making qualitative data more relevant in the clinical setting. LLMs appear to be able to align well with the goals of qualitative interviewing and thematic analysis projects such as: uncovering insights, understandings, and meanings that might not be immediately apparent without in-depth analysis with the depth and context to understand the experiences, attitudes, behaviors, and perceptions of participants. It should be noted that these data do not easily encapsulate into the methods we have outlined here and (for now) still demand a human touch.

References

[1] S. De Paoli, Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach, Social Science Computer Review (2023) 08944393231220483. https://doi.org/10.1177/08944393231220483

[2] T.A. Koleck, C. Dreisbach, P.E. Bourne, S. Bakken, Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review, Journal of the American Medical Informatics Association 26 (2019) 364-379. https://doi.org/10.1093/jamia/ocy173.

[3] M.R. Turchioe, A. Volodarskiy, J. Pathak, D.N. Wright, J.E. Tcheng, D. Slotwiner, Systematic review of current natural language processing methods and applications in cardiology, Heart 108 (2022) 909-916. https://doi.org/10.1136/heartjnl-2021-319769.

[4] F. Pethani, A.G. Dunn, Natural language processing for clinical notes in dentistry: A systematic review, Journal of Biomedical Informatics 138 (2023) 104282. https://doi.org/10.1016/j.jbi.2023.104282.

[5] M.Y. Yan, L.T. Gustad, Ø. Nytrø, Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review, Journal of the American Medical Informatics Association 29 (2022) 559-575. https://doi.org/10.1093/jamia/ocab236.

[6] M. Afshar, A. Phillips, N. Karnik, J. Mueller, D. To, R. Gonzalez, et al., Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation, Journal of the American Medical Informatics Association 26 (2019) 254-261. https://doi.org/10.1093/jamia/ocy166.

[7] B. Mesko, The ChatGPT (Generative Artificial Intelligence) Revolution Has Made Artificial Intelligence Approachable for Medical Professionals, Journal of Medical Internet Research 25 (2023) e48392. https://doi.org/10.2196/48392.

[8] R.K. Garg, V.L. Urs, A.A. Agarwal, S.K. Chaudhary, V. Paliwal, S.K. Kar, Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review, Health Promot Perspect 13 (2023) 183-191. doi: 10.34172/hpp.2023.22. PMID: 37808939; PMCID: PMC10558973.

[9] G. Eysenbach, The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers, JMIR Med Educ 9 (2023) e46885. https://doi.org/10.2196/46885.

[10] J. Clusmann, F.R. Kolbinger, H.S. Muti, Z.I. Carrero, J.-N. Eckardt, N.G. Laleh, et al., The future landscape of large language models in medicine, Commun Med 3 (2023) 1-8. https://doi.org/10.1038/s43856-023-00370-1.

[11] D. Johnson, R. Goodman, J. Patrinely, C. Stone, E. Zimmerman, R. Donald, S. Chang, S. Berkowitz, A. Finn, E. Jahangir, E. Scoville, T. Reese, D. Friedman, J. Bastarache, Y. van der

Heijden, J. Wright, N. Carter, M. Alexander, J. Choe, C. Chastain, J. Zic, S. Horst, I. Turker, R. Agarwal, E. Osmundson, K. Idrees, C. Kieman, C. Padmanabhan, C. Bailey, C. Schlegel, L. Chambless, M. Gibson, T. Osterman, L. Wheless, Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model, Res Sq [Preprint]. 2023 Feb 28:rs.3.rs-2566942. doi: 10.21203/rs.3.rs-2566942/v1. PMID: 36909565; PMCID: PMC10002821.

[12] J.W. Ayers, A. Poliak, M. Dredze, E.C. Leas, Z. Zhu, J.B. Kelley, D.J. Faix, A.M. Goodman, C.A. Longhurst, M. Hogarth, D.M. Smith, Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum, JAMA Intern Med 183 (2023) 589-596. doi: 10.1001/jamainternmed.2023.1838. PMID: 37115527; PMCID: PMC10148230.

[13] E.B. Gordon, A.J. Towbin, P. Wingrove, U. Shafique, B. Haas, A.B. Kitts, J. Feldman, A. Furlan, Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions, J Am Coll Radiol (2023). doi: 10.1016/j.jacr.2023.09.011. Epub ahead of print. PMID: 37863153.

[14] N.M. Barrington, N. Gupta, B. Musmar, D. Doyle, N. Panico, N. Godbole, et al., A Bibliometric Analysis of the Rise of ChatGPT in Medical Research, Medical Sciences 11 (2023) 61. https://doi.org/10.3390/medsci11030061.

[15] V. Braun, V. Clarke, Using thematic analysis in psychology, Qualitative Research in Psychology 3 (2006) 77-101. https://doi.org/10.1191/1478088706qp063oa.

[16] R.H. Tai, L.R. Bentley, X. Xia, J.M. Sitt, S.C. Fankhauser, A.M. Chicas-Mosier, et al., Use of Large Language Models to Aid Analysis of Textual Data, 2023 (2023) 2023.07.17.549361. https://doi.org/10.1101/2023.07.17.549361.

[17] Z. Xiao, X. Yuan, Q.V. Liao, R. Abdelghani, P.-Y. Oudeyer, Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding, 28th International Conference on Intelligent User Interfaces (2023) 75-78. https://doi.org/10.1145/3581754.3584136.

[18] S. Alder, Is ChatGPT HIPAA Compliant?, HIPAA Journal (2023). Available online: https://www.hipaajournal.com/is-chatgpt-hipaa-compliant/ (accessed December 10, 2023).

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. https://doi.org/10.48550/ARXIV.1706.03762.

[20] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C.C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, … T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. https://doi.org/10.48550/ARXIV.2307.09288.

[21] S. Zhao, W. Mathis, Understanding Barriers and Facilitators of Primary Care Use Among Assertive Community Treatment Teams Via Qualitative Analysis of Clients and Clinicians, 2023 (2023) 2023.12.05.23299368. https://doi.org/10.1101/2023.12.05.23299368.

[22] Whisper (2023). Available online: https://openai.com/research/whisper (Accessed December 20, 2023)

[23] Upstage (2023). Available online: https://huggingface.co/upstage/Llama-2-70b-instruct (Accessed December 20, 2023)

[24] ExLlama (2023). Available online: https://github.com/turboderp/exllama (Accessed December 20, 2023).

[25] Text generation web UI (2023). Available online: https://github.com/oobabooga/text-generation-webui (Accessed December 20, 2023).

[26] Preset Arean Results (2023). Available online: https://github.com/oobabooga/oobabooga.github.io/blob/main/arena/results.md (accessed December 10, 2023).

[27] J. Ni, G.H. Ábrego, N. Constant, J. Ma, K.B. Hall, D. Cer, Y. Yang, Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models, 2021. https://doi.org/10.48550/ARXIV.2108.08877.

[28] A. Ziegler, J. Berryman, A developer's guide to prompt engineering and LLMs, The GitHub Blog (2023). Available online: https://github.blog/2023-07-17-prompt-engineering-guide-generative-ai-llms/ (accessed December 10, 2023).

[29] B. Dickson, Optimize your ChatGPT prompts with DeepMind's OPRO technique – TechTalks (2023). Available online: https://bdtechtalks.com/2023/11/20/deepmind-opro-llm-optimization/ (accessed December 10, 2023).

[30] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B.P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, P. Clark, Self-refine: Iterative refinement with self-feedback, 2023. https://doi.org/10.48550/ARXIV.2303.17651.

[31] K. Hebenstreit, R. Praas, L.P. Kiesewetter, M. Samwald, An automatically discovered chain-of-thought prompt generalizes to novel models and datasets, 2023. https://doi.org/10.48550/ARXIV.2305.02897.

[32] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: Language agents with verbal reinforcement learning, 2023. https://doi.org/10.48550/ARXIV.2303.11366.

[33] V. Nair, E. Schumacher, G. Tso, A. Kannan, Dera: Enhancing large language model completions with dialog-enabled resolving agents, 2023. https://doi.org/10.48550/ARXIV.2303.17071.

[34] Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health, 11(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

[35] R.D. Parker, K. Mancini, M.D. Abram, Natural Language Processing Enhanced Qualitative Methods: An Opportunity to Improve Health Outcomes, International Journal of Qualitative Methods 22 (2023) 16094069231214144. https://doi.org/10.1177/16094069231214144.

[36] J.J. Hanna, A.D. Wakene, C.U. Lehmann, R.J. Medford, Assessing Racial and Ethnic Bias in Text Generation for Healthcare-Related Tasks by ChatGPT1, 2023 (2023) 2023.08.28.23294730. https://doi.org/10.1101/2023.08.28.23294730.

[37] M. Sallam, ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns, Healthcare 11 (2023) 887. https://doi.org/10.3390/healthcare11060887.

[38] T. Zack, E. Lehman, M. Suzgun, J.A. Rodriguez, L.A. Celi, J. Gichoya, et al., Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare, 2023 (2023) 2023.07.13.23292577. https://doi.org/10.1101/2023.07.13.23292577.

[39] N. Ito, S. Kadomatsu, M. Fujisawa, K. Fukaguchi, R. Ishizawa, N. Kanda, et al., The Accuracy and Potential Racial and Ethnic Biases of GPT-4 in the Diagnosis and Triage of Health Conditions: Evaluation Study, JMIR Med Educ 9 (2023) e47532. https://doi.org/10.2196/47532.

[40] S. Liu, A.P. Wright, B.L. Patterson, J.P. Wanderer, R.W. Turer, S.D. Nelson, et al., Using AI-generated suggestions from ChatGPT to optimize clinical decision support, J Am Med Inform Assoc 30 (2023) 1237-1245. https://doi.org/10.1093/jamia/ocad072.

[41] A. Jain, J.R. Brooks, C.C. Alford, C.S. Chang, N.M. Mueller, C.A. Umscheid, A.S. Bierman, Awareness of Racial and Ethnic Bias and Potential Solutions to Address Bias With Use of Health Care Algorithms, JAMA Health Forum 4 (2023) e231197. doi: 10.1001/jamahealthforum.2023.1197. PMID: 37266959; PMCID: PMC10238944.

[42] J.K. Paulus, D.M. Kent, Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities, NPJ Digit Med 3 (2020) 99. doi: 10.1038/s41746-020-0304-9. PMID: 32821854; PMCID: PMC7393367.

[43] N. Kordzadeh, M. Ghasemaghaei, Algorithmic bias: review, synthesis, and future research directions, Eur J Inf Syst (2022) 388. doi: 10.1080/0960085x.2021.1927212.

[44] J. Huang, G. Galal, M. Etemadi, M. Vaidyanathan, Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review, JMIR Med Inform 10 (2022) e36388. doi: 10.2196/36388. PMID: 35639450; PMCID: PMC9198828.

[45] H.M. Thompson, B. Sharma, S. Bhalla, R. Boley, C. McCluskey, D. Dligach, M.M. Churpek, N.S. Karnik, M. Afshar, Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across

racial subgroups, J Am Med Inform Assoc 28 (2021) 2393-2403. doi: 10.1093/jamia/ocab148. PMID: 34383925; PMCID: PMC8510285.

[46] F. Li, P. Wu, H.H. Ong, J.F. Peterson, W.Q. Wei, J. Zhao, Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction, J Biomed Inform 138 (2023) 104294. doi: 10.1016/j.jbi.2023.104294. Epub 2023 Jan 24. PMID: 36706849.

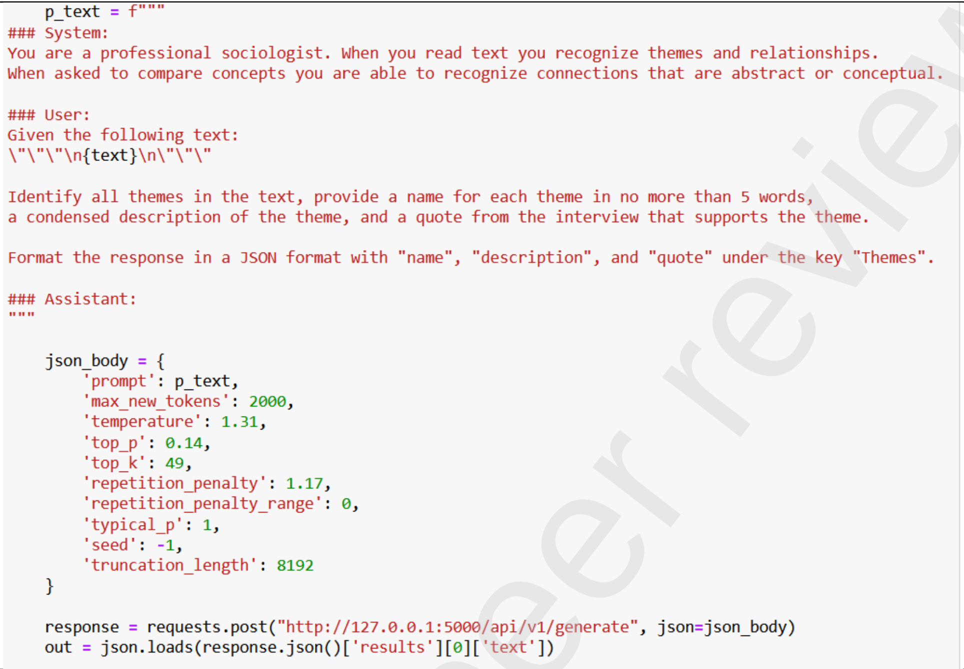
[47] B. Mehra, Bias in Qualitative Research: Voices from an Online Classroom, The Qualitative Report 7 (2002) 1-19. Retrieved from https://nsuworks.nova.edu/tqr/vol7/iss1/2.

[48] Dettmers, T., & Zettlemoyer, L. (2022). The case for 4-bit precision: K-bit Inference Scaling Laws. https://doi.org/10.48550/ARXIV.2212.09720

Supplement

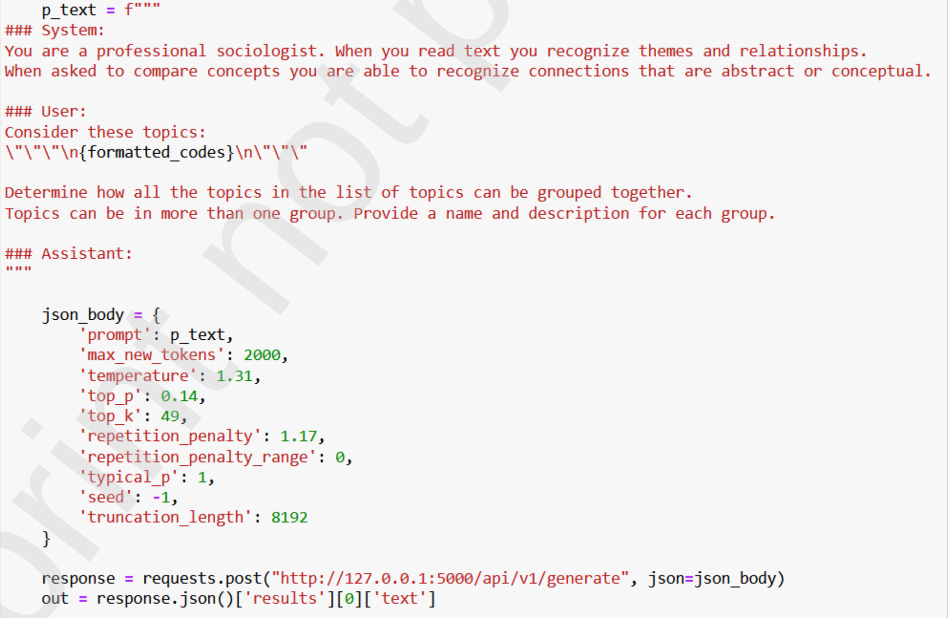
S Table 1. Steps with python code

1. Make Codes – looks at each client and clinician interview, generates and collects list of codes



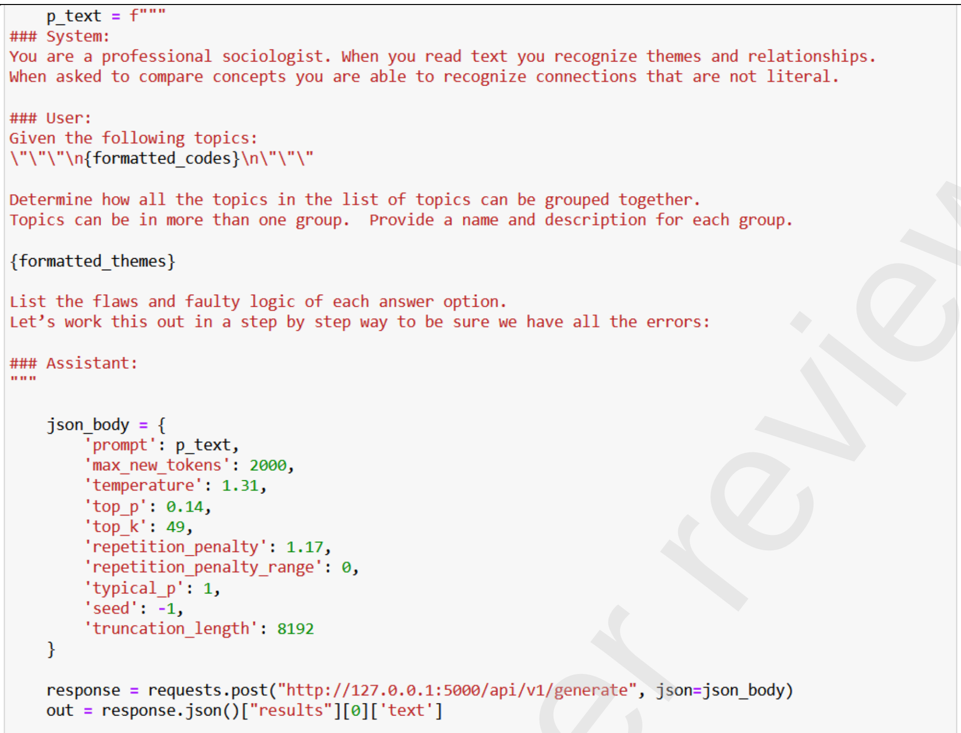
```
    p_text = f"""
### System:
You are a professional sociologist. When you read text you recognize themes and relationships.
When asked to compare concepts you are able to recognize connections that are abstract or conceptual.

### User:
Given the following text:
\"\"\"\n{text}\n\"\"\"

Identify all themes in the text, provide a name for each theme in no more than 5 words,
a condensed description of the theme, and a quote from the interview that supports the theme.

Format the response in a JSON format with "name", "description", and "quote" under the key "Themes".

### Assistant:
"""

    json_body = {
        'prompt': p_text,
        'max_new_tokens': 2000,
        'temperature': 1.31,
        'top_p': 0.14,
        'top_k': 49,
        'repetition_penalty': 1.17,
        'repetition_penalty_range': 0,
        'typical_p': 1,
        'seed': -1,
        'truncation_length': 8192
    }

    response = requests.post("http://127.0.0.1:5000/api/v1/generate", json=json_body)
    out = json.loads(response.json()['results'][0]['text'])
```

2. Make Themes – takes list of codes and generates themes – this is run for 3 iterations.



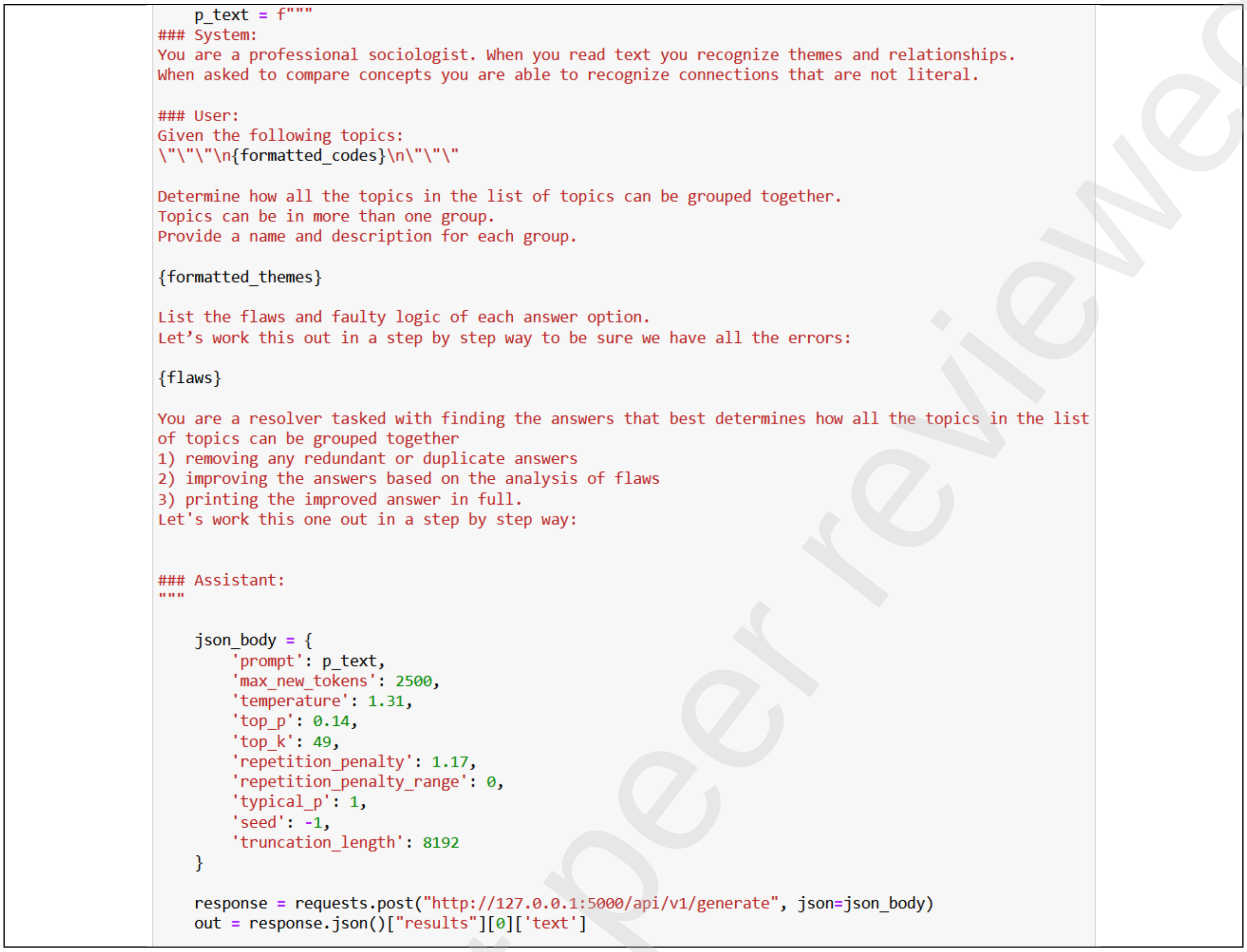
```
    p_text = f"""
### System:
You are a professional sociologist. When you read text you recognize themes and relationships.
When asked to compare concepts you are able to recognize connections that are abstract or conceptual.

### User:
Consider these topics:
\"\"\"\n{formatted_codes}\n\"\"\"

Determine how all the topics in the list of topics can be grouped together.
Topics can be in more than one group. Provide a name and description for each group.

### Assistant:
"""

    json_body = {
        'prompt': p_text,
        'max_new_tokens': 2000,
        'temperature': 1.31,
        'top_p': 0.14,
        'top_k': 49,
        'repetition_penalty': 1.17,
        'repetition_penalty_range': 0,
        'typical_p': 1,
        'seed': -1,
        'truncation_length': 8192
    }

    response = requests.post("http://127.0.0.1:5000/api/v1/generate", json=json_body)
    out = response.json()['results'][0]['text']
```

3. Find flaws – ask the LLM to evaluate the generated themes for flaws and faulty logic



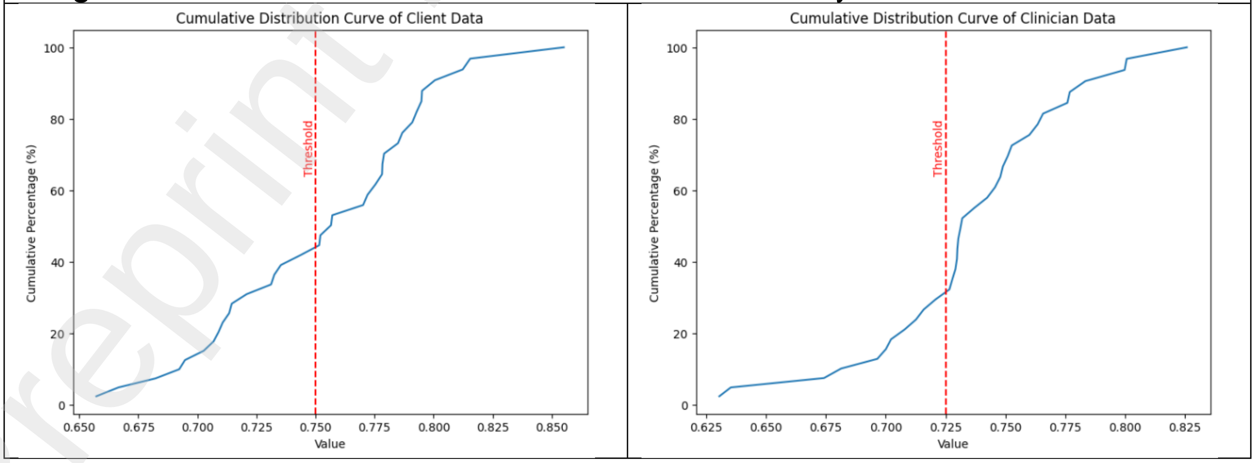
```
    p_text = f"""
### System:
You are a professional sociologist. When you read text you recognize themes and relationships.
When asked to compare concepts you are able to recognize connections that are not literal.

### User:
Given the following topics:
\"\"\"\n{formatted_codes}\n\"\"\"

Determine how all the topics in the list of topics can be grouped together.
Topics can be in more than one group.  Provide a name and description for each group.

{formatted_themes}

List the flaws and faulty logic of each answer option.
Let’s work this out in a step by step way to be sure we have all the errors:

### Assistant:
"""

    json_body = {
        'prompt': p_text,
        'max_new_tokens': 2000,
        'temperature': 1.31,
        'top_p': 0.14,
        'top_k': 49,
        'repetition_penalty': 1.17,
        'repetition_penalty_range': 0,
        'typical_p': 1,
        'seed': -1,
        'truncation_length': 8192
    }

    response = requests.post("http://127.0.0.1:5000/api/v1/generate", json=json_body)
    out = response.json()["results"][0]['text']
```

4. Resolve – ask the LLM to incorporate this feedback and produce improved themes



```
    p_text = f"""
### System:
You are a professional sociologist. When you read text you recognize themes and relationships.
When asked to compare concepts you are able to recognize connections that are not literal.

### User:
Given the following topics:
\"\"\"\n{formatted_codes}\n\"\"\"

Determine how all the topics in the list of topics can be grouped together.
Topics can be in more than one group.
Provide a name and description for each group.

{formatted_themes}

List the flaws and faulty logic of each answer option.
Let’s work this out in a step by step way to be sure we have all the errors:

{flaws}

You are a resolver tasked with finding the answers that best determines how all the topics in the list
of topics can be grouped together
1) removing any redundant or duplicate answers
2) improving the answers based on the analysis of flaws
3) printing the improved answer in full.
Let's work this one out in a step by step way:


### Assistant:
"""

    json_body = {
        'prompt': p_text,
        'max_new_tokens': 2500,
        'temperature': 1.31,
        'top_p': 0.14,
        'top_k': 49,
        'repetition_penalty': 1.17,
        'repetition_penalty_range': 0,
        'typical_p': 1,
        'seed': -1,
        'truncation_length': 8192
    }

    response = requests.post("http://127.0.0.1:5000/api/v1/generate", json=json_body)
    out = response.json()["results"][0]['text']
```

S Fig 1 – Threshold determination for Sentence-T5-xxl similarity binarization


Cumulative Distribution Curve of Client Data
Cumulative Percentage (%)
Value
Threshold
Cumulative Distribution Curve of Clinician Data
Cumulative Percentage (%)
Value
Threshold