RACER: An LLM-powered Methodology for Scalable Analysis of Semi-structured Mental Health Interviews

Satpreet H Singh[1*]
Baylor College of Medicine
satpreetsingh@gmail.com

Kevin Jiang[1]
Baylor College of Medicine,
(Houston, TX)

Kanchan Bhasin[2]
Rice University
(Houston, TX)

Ashutosh Sabharwal[2†]
Rice University

Nidal Moukaddam[1†]
Baylor College of Medicine

Ankit B Patel[1,2†]
Baylor College of Medicine
Rice University

Abstract

Semi-structured interviews (SSIs) are a commonly employed data-collection method in healthcare research, offering in-depth qualitative insights into subject experiences. Despite their value, manual analysis of SSIs is notoriously time-consuming and labor-intensive, in part due to the difficulty of extracting and categorizing emotional responses, and challenges in scaling human evaluation for large populations. In this study, we develop RACER, a Large Language Model (LLM) based expert-guided automated pipeline that efficiently converts raw interview transcripts into insightful domain-relevant themes and sub-themes. We used RACER to analyze SSIs conducted with 93 healthcare professionals and trainees to assess the broad personal and professional mental health impacts of the COVID-19 crisis. RACER achieves moderately high agreement with two human evaluators (72%), which approaches the human inter-rater agreement (77%). Interestingly, LLMs and humans struggle with similar content involving nuanced emotional, ambivalent/dialectical, and psychological statements. Our study highlights the opportunities and challenges in using LLMs to improve research efficiency and opens new avenues for scalable analysis of SSIs in healthcare research.

1 Introduction

Semi-structured interviews (SSIs) are a widely used qualitative research method in healthcare research that provide an in-depth understanding of subjects' experiences in their own words (Adams, 2010). SSIs require interviewers to ask pre-specified 'root' questions, along with the option to ask follow-up questions to gain clarity on the interviewee's responses. This flexibility is a key characteristic of SSIs, allowing for a more dynamic and responsive data collection process, especially in areas where exploratory forays are needed. The adaptability of SSIs is particularly beneficial in exploring complex or sensitive topics such as mental health. SSIs allow rapport building between interviewer and subject and facilitate candid responses on sensitive matters. The open-ended nature of follow-up questions gives subjects the freedom to reflect on experiences and articulate thoughts without judgement. This helps reveal the nuances, contradictions, and diversity of perspectives that traditional fixed quantitative surveys may overlook. However, the traditional manual analysis of these interviews is a time-consuming and resource-intensive process. The advent of Large Language Models (LLMs), such as GPT-4 (Lee et al., 2023b,a,e), offers a novel and efficient method to extract and interpret data from such text corpora. Yet, the validity of LLMs in analyzing emotional states may be limited in circumstances where participants express multiple emotions or conflicting (dialectical) states.

As a case-study, we leveraged data from SSIs, conducted during the peak of the COVID-19 crisis in 2020, to understand the mental well-being of 93 healthcare professionals and trainees. The COVID-19 pandemic brought to the forefront profound personal and professional challenges experienced by healthcare workers. Fear of infecting family members, grief over patient deaths, moral dilemmas in resource allocation, and anxieties about professional preparedness collectively introduced a heightened level of psychological complexity and stress in the lives of healthcare professionals. The stigma surrounding the pursuit of mental health support exacerbated these challenges, leaving healthcare workers hesitant to openly discuss their difficulties or seek assistance.

In this paper, we developed **RACER**, an expert-guided automated pipeline that **R**etrieved responses to about 40 questions per SSI, **A**ggregated responses to each question across all subjects, **C**lustered these responses for each question into insightful domain-relevant **E**xpert-guided themes (Lee et al., 2023c), and finally **R**e-clustered

Proceedings of the 1st Workshop on NLP for Science (NLP4Science), pages 73–98
November 16, 2024 ©2024 Association for Computational Linguistics

responses to produce a robust result. Human evaluation on a subset of the total population revealed moderately high agreement (McHugh, 2012) between humans and RACER outputs, and similarities between inter-human disagreement and human-machine disagreement. We summarize our findings from applying RACER to our SSI-survey on the experiences of healthcare professionals and trainees during COVID-19, to reveal the power of this approach. Our results demonstrate both the capabilities and the limitations leveraging LLMs to efficiently process and extract insights from a large corpus of SSIs.

Related Work

Our research is related to a growing body of research that applies state-of-the-art and open-source LLMs to medical (Clusmann et al., 2023; Shah et al., 2023a) and psychological text corpora (Stade et al., 2024), with the most common and related applications being in mental health chatbots (Lee et al., 2023a) and medical evidence summarization and documentation (Tang et al., 2023a; Wornow et al., 2023a; Shah et al., 2023b). This literature reports broad improvements in performance over previous methods using classic Natural Language Processing (NLP) techniques in such domains (Raveau et al., 2023) Our research is most similar to very recent work assessing the use of LLMs in psychiatric mental health assessment (Kjell et al., 2024) and thematic analysis more broadly (Dai et al., 2023; Lee et al., 2023d; Stefano De Paoli, 2023), where the authors produce one-off examples of LLMs applied to specific use-cases replacing traditional research methods. In contrast, we present an expert-guided, reliable, and scalable methodology for SSI analysis, and an end-to-end case study applying our methodology to a real-world dataset, to demonstrate the efficacy of our methods for mental-health and burnout related SSIs. Furthermore, our analyses reveal intriguing similarities between inter-human disagreement and the self-consistency of LLM outputs.

2 Results

Recruitment and interview of a diverse sample of healthcare professionals and trainees

Healthcare professionals and trainees across different specialties and career stages were recruited via snowball sampling method (Goodman, 1961),

Characteristic	**Percentage**
Gender	
Male	54.84%
Female	45.16%
Age Group	
22-33 years	39.78%
34-45 years	32.26%
46-60 years	16.13%
61+ years	5.38%
Unclear/Excluded	6.45%
Healthcare Professional/Student Type (non-exclusive membership)	
Physicians	54.84%
Medical Students	21.51%
Nurses	8.60%
Residents	7.53%
Other Professionals	12.90%
Unclear/Excluded	1.08%
Location	
Houston, Texas	44.09%
Other Texas	21.50%
Florida	10.75%
Mid-West US	13.98%
Other US	5.38%
Unclear/Excluded	4.30%
Marital Status	
Not married	41.94%
Married	52.69%
Unclear/Excluded	5.37%
Have Kids?	
Yes	51.61%
No	45.16%
Unclear/Excluded	3.23%
Specialty Area (non-exclusive membership)	
Emergency Medicine	26.88%
Psychiatry	16.13%
Pulmonary Critical Care	16.13%
Internal Medicine	11.83%
Neurology/Neurocritical Care	5.38%
Surgery/ER	5.38%
Pediatrics	5.38%
Other Specialties	17.22%
Unclear/Excluded	2.15%
Years of Practice (Non-students)	
Under 15 Years	71.23%
15-30 Years	20.55%
Over 30 Years	5.48%
Unclear/Excluded	2.74%

Table 1: Demographic Characteristics of the Study Population. Note that some categories are non-exclusive. e.g. practicing faculty are categorized under both Physicians and Other Professionals.

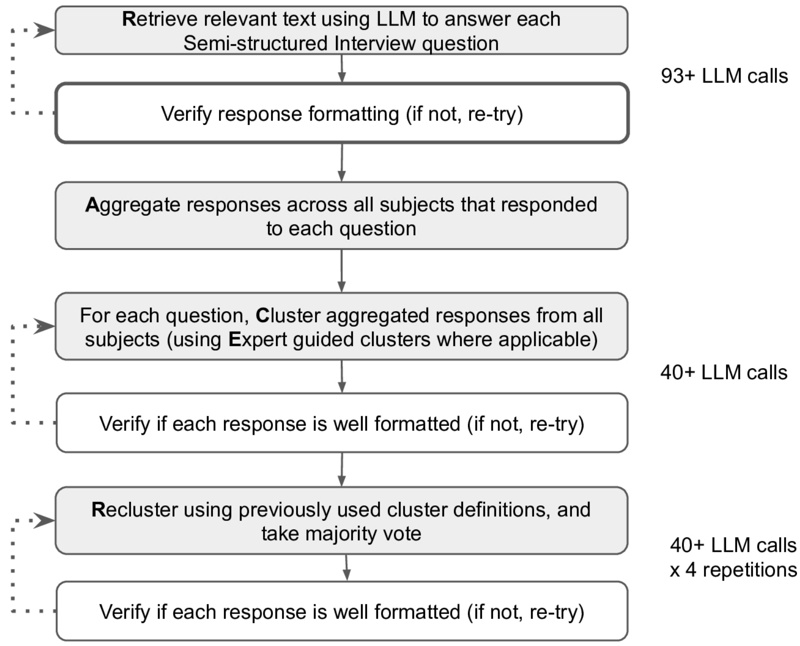

Figure 1: **Stages of the RACER (Retrieve, Aggregate, Cluster with Expert guidance, and Re-cluster) Semi-Structured Interview (SSI) processing pipeline:** First, **R**etrieve relevant responses to each SSI question. **A**ggregate responses across subjects before **C**lustering them into themes (and subthemes) defined by **E**xperts. To assess robustness, **R**e-cluster multiple times and make assignments by majority vote. The pipeline efficiently and robustly converts SSI text into meaningful themes.

described as follows. The investigators asked colleagues if they knew of anyone willing to participate in interviews about their COVID-19 experiences. Announcements were also posted online and through professional networks. Participation was voluntary with no compensation provided. Approval was obtained from the Baylor College of Medicine (Houston, TX) Institutional Review Board. The interviews were performed by a team of two research coordinators with healthcare backgrounds, and a third-year medical student, under the supervision of the investigators.

The study population of healthcare professionals and trainees consisted of 93 subjects (51 male, 42 female) with diverse demographics (Table 1). Subjects were from 22 years to over 61 years in age, and were located predominantly in Texas. Over half were married and had children. Most subjects had no care-taking responsibilities in addition to child-care. Professionally, the sample included physicians, medical students, nurses, residents and other healthcare professionals. Subjects trained at multiple institutions, with prominent representation from Baylor College of Medicine and University of Texas systems. Various specialties were represented in the cohort, with emergency medicine, psychiatry and pulmonary/critical care among the most common.

SSIs were conducted over videoconferencing using a standard template consisting of a total of 41 questions, including four questions that were only asked to students, and seven questions that were asked to only non-students. Questions were either *factual*, concerning demographics and personal and professional background, or *open ended*, where interviewees were asked to talk about their experiences during the COVID-19 pandemic, focusing on their exposure to the virus, work impacts, emotional responses, future outlooks, and coping strategies. Interviewees discussed how they had practiced in high-risk areas, their concerns for personal and family safety, and modifications made to their routines. They also reflected on the physical toll the crisis had taken. The impact on their work included changes in working hours, shifts in patient care quality, and altered management approaches. Emotional and psychological questions revealed how the crisis affected them emotionally, the level of support they received, family dynamics, and changes in burnout levels. Looking ahead, they pondered the crisis's short-term and long-term impacts on their careers and specialty choices. Finally, they shared their openness to seeking help for burnout or mental overwhelm and identified potential obstacles in obtaining this help. Students were not asked clinical-practice related questions, and were instead asked about how their training was being affected by pandemic-related changes. Interviews lasted on average 26.7 +/- 8.9 s.d. minutes. When transcribed from raw interview audio into text transcripts (using Otter.AI(Otter.ai, 2023)), were on average 4044.30 +/- 1348.34 s.d. words long.

RACER extracts relevant interviewee responses and robustly clusters them

We developed an LLM-based automated pipeline called **RACER** (Figure 1) that converts a corpus of text SSI transcripts into insightful themes per interview question. RACER, consists of four stages, **R**etrieve, **A**ggregate, **C**luster with **E**xpert guidance, and **R**ecluster:

Retrieve: We first structured interview transcripts by using an LLM (OpenAI's GPT-4(Lee et al., 2023b)) to *retrieve* relevant SSI text in response to each of the questions in the interview template. (See Appendix A for LLM prompt details) To avoid LLM 'hallucinations' (Tonmoy et al., 2024), we asked the LLM to provide 'evidence' in the form of text quoted verbatim from the transcript, to back up its response to each question. LLM outputs missing either answers or backing evidence to any question were automatically detected and re-run.

Aggregate: For each question, we then aggregated the retrieved responses across all subjects who were asked that question.

Cluster with Expert guidance: We then asked the LLM to *semantically cluster* the responses into primary and secondary clusters ('themes' and 'sub-themes'). For most questions, we provided the LLM expert-guidance in the form of primary-cluster definitions. These definitions were derived through a combination of theoretical foundations from burnout literature and practical insights from ongoing research during the COVID-19 pandemic (Moukaddam et al., 2022; Innstrand, 2022; Edú-Valsania et al., 2022). The primary clusters were selected on the basis of well-established symptom categories of burnout, such as emotional exhaustion, depersonalization/detachment, and cynicism, as well as factors exacerbated by the pandemic, like involvement with COVID-19 patients, fear of spreading the disease, and COVID-19 induced stress. This process involved expert review of early LLM experiments, where we observed that the LLM's autonomous clustering could be too variable or too fine-grained for statistical analysis. We then designed a few primary clusters per question such that clusters were mutually exclusive and collectively exhaustive.

For questions where primary clusters were not derived from expert-guidance, we allowed the LLM to autonomously discover primary clusters. In these cases, the LLM's discovered clusters were reviewed by experts to ensure they were meaningful and useful for subsequent analysis.

The LLM discovered secondary clusters (or sub-themes) automatically. Expert-provided cluster definitions were always mutually exclusive and collectively exhaustive, while those discovered by the LLM were not constrained to be so. Similar to before, invalid LLM responses, e.g. those missing cluster assignments for any subjects, were automatically re-run.

This approach thus leveraged the strengths of both expert knowledge and LLM capabilities. See Supplementary Tables 2 and 3 for expert-guided and LLM-discovered primary clusters respectively.

Re-Cluster: Leveraging the probabilistic nature of LLMs, we assessed the *robustness* of the clustering process by re-running it four more times, employing the same cluster definitions and validation criteria as in the initial step. We used a majority vote over 5 runs to assign subjects to clusters, to get robust cluster assignments for all downstream processing. The number of votes (3, 4 or 5 out of total 5 LLM calls) additionally provided a synthetic measure of LLM *self-consistency* (Kompa et al., 2021; Tanneru et al., 2023) that we have quantified as a 'self-concordance score'. Only a very small fraction of subject-question pairs (12 out of 3342, 0.36%) had no 'self-concordant' cluster assignments after applying the majority voting process.

All together, we found that RACER was able to take unstructured transcriptions and extract relevant and insightful, clustered responses in a robust manner for downstream human analysis.

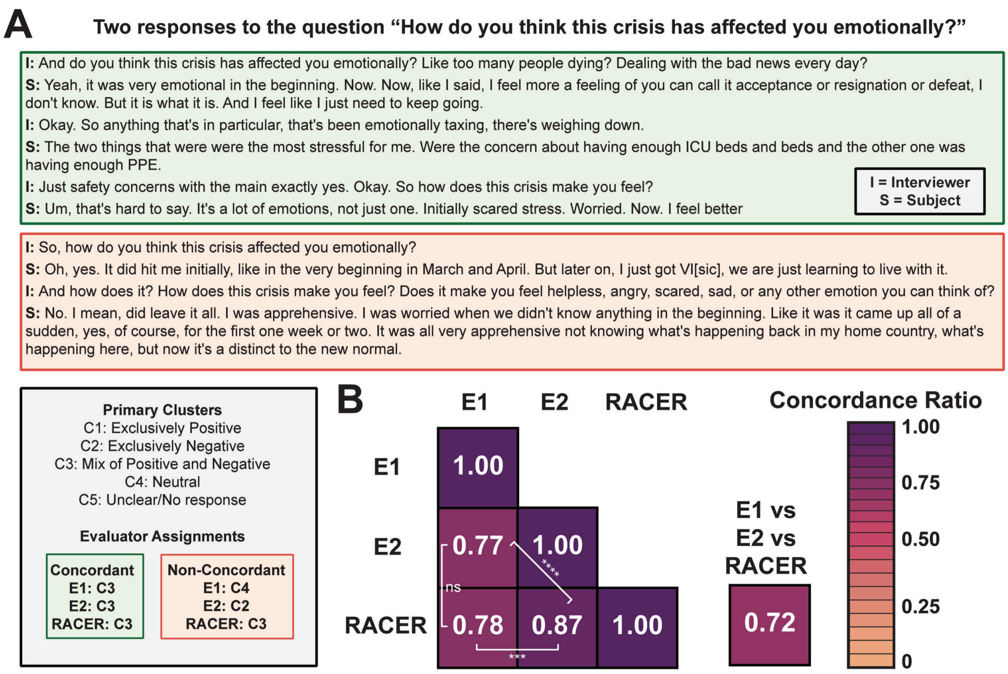

Figure 2: **Human-RACER approaches resembles human-human disagreement:** (A) Transcript segments from two different subjects being asked "How do you think this [COVID-19] crisis has affected you emotionally?". Responses were evaluated as either all concordant or all non-concordant between both evaluators and RACER, demonstrating the ambiguity that exists in parsing free responses. (B) The concordance ratio calculated between evaluator pairs, and between RACER and both evaluators simultaneously. Chi-squared test with Yates continuity correction between the three different evaluator pairings showed human evaluator concordance did not differ from evaluator one's concordance with RACER. * $p < 0.5$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

Human-machine disagreement approaches inter-human disagreement

To validate the output of running RACER on our SSI dataset, two human evaluators cross-checked the resulting cluster assignments for 20 randomly-selected subjects across 28 open ended questions (See Figure 2A for an example). Using the same cluster definitions as were previously used by RACER, each human evaluator (E1 and E2) independently read the raw transcript file and assigned each subject's answers to the primary clusters. Evaluator cluster assignments were then compared to RACER's robust cluster assignments. To quantify agreement, we defined a *concordance score* and a *concordance ratio* as follows: If the clusters for a given subject-question pair matched exactly (for mutually exclusive clusters), or matched partially (for mutually non-exclusive clusters) they were assigned a concordance score of 1. Conversely, mismatch was assigned a concordance score of 0. The overall concordance ratio is the proportion of matched subject-question pairs between evaluators.

We observed a concordance ratio of 78% (E1) and 87% (E2) between each of the human evaluators and RACER, and a 77% (E1-E2) inter-rater concordance ratio (Figure 2B). When the two human evaluators and RACER were compared simultaneously, there was only a small decrease in the concordance ratio (72%), indicating that across the majority of subject-question pairings, cluster assignments produced by humans and RACER were all in agreement. (See Appendix A for additional details)

Machine "confusion" resembles human confusion

We examined the self-concordance produced by RACER per subject-question pair to see how it might affect the subject-question pair's inter-rater concordance (Figure 3).

Amongst the 443 subject-question pair sample evaluated by humans, 392 (87.7%) had a self-concordance of 1 (5 of 5 repeated primary clusters), which was not different proportionally to that of the whole population: 88.2% (1852 of 2099 subject-question pairs), thus RACER's self-

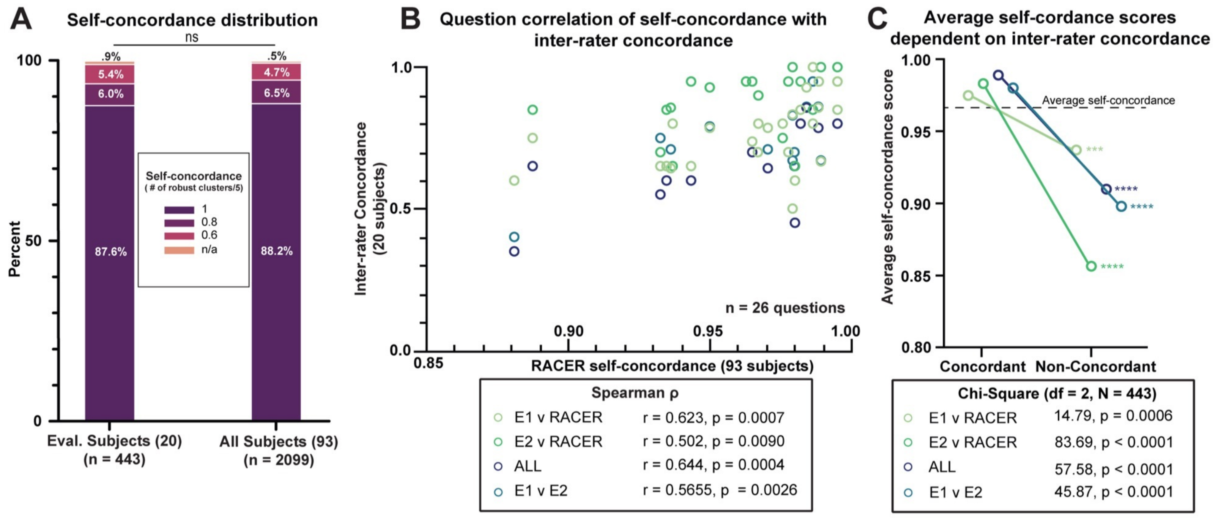

Figure 3: **RACER "self-concordance" correlates with inter-evaluator concordance and reveals areas of human disagreement:** (A) Distribution of the proportion of subject-question pair self-concordance, calculated as the fraction of identical primary cluster assignments across five runs. The self-concordance for the subject-question pairs reviewed by human evaluators (20 subjects) were not significantly different from those for all subject-question pairs (93 subjects), as determined by a Chi-squared test. (B) Average RACER self-concordance for each question (n = 93) show a significant correlation with the concordance between evaluator pairs for the same questions (n = 20), using Spearman Rank correlation. (C) Comparison of RACER self-concordance within concordant versus non-concordant subject-question pairs between human evaluators. The Chi-squared test indicates significant differences in the distribution of self-concordance between these groups. Correlation significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

concordance across the evaluated 20 subjects was representative of its general performance in primary cluster assignment. When RACER's average self-concordance across all subjects for a given question was correlated with the question's inter-rater concordance from the 20 human evaluated subjects, there was a significant and positive correlation between self- and inter-rater concordance. Additionally, we observed that the self-concordance of subject-question pairs that had inter-rater concordance were higher than those that did not regardless of the rater pair compared: human evaluators or RACER.

Interestingly, when we juxtaposed RACER self-concordance against human-human inter-rater concordance, we observed that RACER self-concordance was lower when humans were non-concordant. This suggests that areas where RACER was less self-concordant or 'confused' were also areas where human evaluators tended to disagree. Thus the RACER self-concordance generated via repeated clustering could also serve as an indicator of ambiguity or difficulty of understanding the semi-structured interview and parsing human free-responses.

3 Insights using RACER on healthcare worker experience during COVID-19

We summarize RACER-derived insights from analyzing our 93-subject SSI corpus in Appendix B.

4 Discussion

Summary

Our study demonstrates the utility of RACER for efficiently analyzing semi-structured interviews (SSIs), particularly those exploring complex mental health topics within the healthcare domain. We introduce a novel approach by employing RACER to analyze emotions and psychological behaviors, opening new possibilities for exploration in mental health. By providing expert-guided constraints and using automated response validation steps, RACER accurately extracts and robustly clusters relevant responses from interview transcripts. Automating these laborious manual tasks significantly enhances the scalability of SSI analysis. The inter-rater agreement between LLM-assigned clusters and human expert clusters further bolsters our claims. The automated pipeline achieved moderately high concordance compared with manual evaluation by human annotators. The overall concordance ratio of

0.72 for RACER versus both human evaluators approaches the 0.77 concordance ratio between the two human evaluators.

The robust semantically clustered summary of the SSI corpus is useful to researchers in multiple ways: Clusters reveal common themes and experiences across the population, allowing identification of major issues and concerns. The quantitative breakdowns by cluster provide an overview of the distribution of different sentiments and impacts. These could potentially be used for clinical applications such as early burnout detection, and operational improvements through triage of targeted interventions and support. Since semi-structured textual data has been converted to structured data, comparisons between subgroups (e.g. by demographics or professions) can be used to identify disparities and facilitate equitable allocation of resources. RACER also enables large-scale, multi-site analyses of SSIs by providing a consistent and reproducible methodology for extracting insights from free-text responses, reducing inconsistencies arising from inherent variability between human evaluators across different sites.

Limitations

Our findings reveal both the promises and current pitfalls of LLMs for SSI analysis. We found that when RACER struggled with robust clustering, both humans and machines were more likely to be non-concordant, suggesting shared limitations in handling complex emotions or psychologically nuanced statements (Boag et al., 2021) or ambiguity of the underlying SSI. This underscores the indispensable role of human expertise in reviewing and interpreting LLM outputs, where RACER's self-concordance can guide expert scrutiny.

While RACER provided evidence in the form of quoting relevant interview text to support its response in the Retrieval step, the underlying methodology remains opaque. In contrast, human evaluators were able to describe their techniques, even if subjective. For instance, humans considered different amounts of contextual information outside the question scope, and inferred subject intentions to varying degrees, i.e. whether the subject needed to explicitly say certain phrases, or if they could be inferred from previous statements or knowledge of the subject matter. An LLM's ability to consider large amount of contextual information can be a double-edged sword; beneficial if relevant information appears elsewhere in the transcript, but misleading if the research is indeed directed towards a narrow window of text around the question.

We demonstrated that LLMs can help discover knowledge by automatically extracting themes and topics from subject responses. However, good performance requires clear, mutually exclusive category definitions. We found it highly useful to involve domain experts early to precisely define mutually exclusive thematic clusters. For certain questions, where succinct mutually exclusive categorization was not possible, we chose to use LLM-discovered clusters. However, validation of such non-exclusive categorization is challenging. Our results showed higher LLM accuracy and inter-rater agreement for questions with non-overlapping expert-defined clusters versus those allowing multiple clusters.

Additionally, human evaluators exhibited biases, such as default cluster tendencies requiring countering evidence (e.g. starting from a default of 'no' and requiring evidence to switch to a 'yes', or vice versa). Thus, expert human analysis also demonstrates cognitive variability and individual biases. Rather than definitive classifications, both human and machine outputs should be considered informed yet inherently biased perspectives on complex qualitative responses (Atari et al., 2024). Thus, in the future, clearly delineating the parameters of evaluations with humans and RACER may improve concordance.

While RACER's cluster assignments may deviate slightly from human reviewers, RACER was internally consistent and demonstrated high clustering repeatability for most questions. Furthermore, unlike humans, RACER was able to efficiently process an extensive dataset of 93 subjects and can scale to significantly larger data set sizes that would otherwise be infeasible for human evaluators to handle.

Future work

For researchers undertaking projects in this emerging domain, both optimism and caution are warranted (Badal et al., 2023; Dash et al.; Chiu et al., 2024; Tang et al., 2023b; Wornow et al., 2023b; Shah et al., 2023b). With appropriate constraints and validation, LLMs can accelerate knowledge extraction from SSIs. We implemented safeguards against hallucination risks like requiring verbatim textual evidence for an answer, which constrained the LLM to mostly avoid fabricating content. While this is already an area of active research, the possibility of a few false positives remains and needs to be accounted for in downstream use.

While evaluation of LLM outputs through comparison to multiple human raters is helpful, inter-rater agreement must also be looked at to assess inherent ambiguity. To further improve performance, we recommend specialized training for both SSI interviewers and human evaluators.

We found it useful to generate an ensemble of LLM clustering outputs from repeated runs, and used it to extract robust cluster assignments and to get a measure of model uncertainty. Future work exploring this direction could produce useful methods that help build trust in LLM-assisted analyses and inform human-in-the-loop processes for high-stakes applications (Bienefeld et al., 2023).

References

Eike Adams. 2010. The joys and challenges of semi-structured interviewing. *Community Practitioner*, 83(7):18–22.

Mohammad Atari, Mona J. Xue, Peter S. Park, Damián Blasi, and Joseph Henrich. 2024. Which Humans? Publisher: OSF.

Kimberly Badal, Carmen M. Lee, and Laura J. Esserman. 2023. Guiding principles for the responsible development of artificial intelligence tools for healthcare. *Communications Medicine*, 3(1):47.

Nadine Bienefeld, Jens Michael Boss, Rahel Lüthy, Dominique Brodbeck, Jan Azzati, Mirco Blaser, Jan Willms, and Emanuela Keller. 2023. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. *npj Digital Medicine*, 6(1):94.

William Boag, Olga Kovaleva, Thomas H. McCoy, Anna Rumshisky, Peter Szolovits, and Roy H. Perlis. 2021. Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Translational Psychiatry*, 11(1):1–6. Number: 1 Publisher: Nature Publishing Group.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A Computational Framework for Behavioral Assessment of LLM Therapists. *arXiv preprint*. ArXiv:2401.00820 [cs].

Jan Clusmann, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P. Veldhuizen, Sophia J. Wagner, and Jakob Nikolas Kather. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. Llm-in-the-loop: Leveraging large language model for thematic analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9993–10001.

Debadutta Dash, Rahul Thapa, Juan M Banda, Akshay Swaminathan, Mehr Kashyap, Nikesh Kotecha, Jonathan H Chen, Saurabh Gombar, Lance Downing, Rachel Pedreira, Ethan Goh, Angel Arnaout, Garret K Morris, Matthew P Lungren, Eric Horvitz, and Nigam H Shah. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery.

Sergio Edú-Valsania, Ana Laguía, and Juan A Moriano. 2022. Burnout: A review of theory and measurement. *International journal of environmental research and public health*, 19(3):1780.

Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics*, pages 148–170.

Siw Tone Innstrand. 2022. Burnout among health care professionals during covid-19. *International Journal of Environmental Research and Public Health*, 19(18):11807.

Oscar N.E. Kjell, Katarina Kjell, and H. Andrew Schwartz. 2024. Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, 333:115667.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023a. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

Peter Lee, Carey Goldberg, and Isaac Kohane. 2023b. *The AI revolution in medicine: GPT-4 and beyond*. Pearson.

V. Vien Lee, Stephanie C. C. van der Lubbe, Lay Hoon Goh, and Jose M. Valderas. 2023c. Harnessing ChatGPT for thematic analysis: Are we ready? *arXiv preprint*. ArXiv:2310.14545 [cs].

V. Vien Lee, Stephanie C. C. van der Lubbe, Lay Hoon Goh, and Jose M. Valderas. 2023d. Harnessing ChatGPT for thematic analysis: Are we ready? *Preprint*, arxiv:2310.14545.

V Vien Lee, Stephanie CC van der Lubbe, Lay Hoon Goh, and Jose M Valderas. 2023e. Harnessing chatgpt for thematic analysis: Are we ready? *arXiv preprint arXiv:2310.14545*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Nidal Moukaddam, Vishwanath Saragadam, Mahsan Abbasi, Matt Barnett, Anil Kumar Vadathya, Ashok Veeraraghavan, and Ashutosh Sabharwal. 2022. Evolution of mood symptomatology through the covid-19 pandemic: findings from the covidsense longitudinal study. *Cureus*, 14(10).

Otter.ai. 2023. Otter.ai: Voice meeting notes transcription service. Accessed: 2023-12-10.

María P. Raveau, Julián I. Goñi, José F. Rodríguez, Isidora Paiva-Mack, Fernanda Barriga, María P. Hermosilla, Claudio Fuentes-Bravo, and Susana Eyheramendy. 2023. Natural language processing analysis of the psychosocial stressors of mental health disorders during the pandemic. *npj Mental Health Research*, 2(1):17.

Nigam H. Shah, David Entwistle, and Michael A. Pfeffer. 2023a. Creation and Adoption of Large Language Models in Medicine. *JAMA*, 330(9):866.

Nigam H. Shah, David Entwistle, and Michael A. Pfeffer. 2023b. Creation and Adoption of Large Language Models in Medicine. *JAMA*, 330(9):866.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1):12.

Stefano De Paoli. 2023. 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*.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023a. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023b. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2023. Quantifying uncertainty in natural language explanations of large language models. *arXiv preprint arXiv:2311.03533*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. 2023a. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. 2023b. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135.

APPENDICES

A Methods

Semi-structured interviews

Study was approved by the Baylor College of Medicine (Houston, TX) Institutional Review Board [Protocol H-47690]. Consent was obtained by reading the consent text and documenting approval to participate, as the interviews were virtual. All interviewees were adults. Interviewers were provided with a standard template to guide their discussions. The subjects were all healthcare professionals or trainees, including physicians, nurses, and medical students. The interviews followed a semi-structured format, where the interviewers were instructed to cover a previously decided list of questions, and were allowed to ask exploration questions if the 'root' question was not answered. The questions covered in the SSIs are listed in Appendices 2 and 3. Raw audio and video interview files were transcribed into text format using the Otter.AI transcription service (Otter.ai, 2023). Out of 100 interviews conducted, 7 were compromised due to data-corruption/loss issues, providing a total of 93 transcriptions for further processing. Voice to text transcription was carried out using Otter.AI(Otter.ai, 2023), which attempts to perform automated speaker diarization, but does not do so perfectly. To the best of our knowledge, this shortcoming did not seem to influence the subsequent processing steps.

RACER

We used the OpenAI GPT-4 LLM for all our work, except for prompts which exceeded GPT-4's limits, where we used GPT-4-32k.

Retrieval: In this step, the model was tasked with retrieving relevant responses for each question from a predefined list of questions (listed in Appendix E) from the transcript. The prompt for the LLM consisted of instructions and a template consisting of the aforementioned list of questions and what format each question's response should be in, followed by the entire SSI transcript. The full prompt is detailed in Appendix E.

LLM Response Validation for Retrieval: By asking the LLM to respond in a structured format, we could partially automate the process of verifying the LLM's response. The LLM is called once for each subject, and then the response is parsed using the Python Pandas library. The LLM's response is marked invalid if it is ill-formatted (not parsable in tab-separated-values format) or incomplete (wrong number of rows, i.e. questions, or columns, i.e. incomplete response). The LLM is called again on invalid responses till the LLM returns a valid response. We found that at most 4-5 (5%-6%) subjects would have invalid responses in the first attempt, and in total, we were making about 10% additional calls to get valid responses for all subjects. The most common issues were that the LLM would sometimes be incomplete (skip questions, end output before final question) and sometimes use the specified tab-delimiter incorrectly.

Cluster with Expert guidance: In this step, we employed a semantic clustering approach which grouped responses based on the underlying themes or sentiments ("semantic clusters") they conveyed.

Expert Guidance: In preliminary explorations, we found that the LLM is able to automatically generate interesting semantic clusters from a list of the subjects' responses without additional human guidance. We observed that these clusters could change between subsequent LLM calls, could be mutually non-exclusive (subjects could belong to multiple clusters), and could be too fine-grained for statistical analysis. However, in many cases (29 out of $\approx$40 questions, see Appendix C), we felt like it was important to exercise more control over the LLM's response to improve response robustness, to facilitate statistical analysis and for easier human evaluation. So, we provided expert guidance in the form of a list of primary clusters or "themes" (defined on a per-question basis), which were included in the prompt using a template (detailed in Appendix F). Secondary clusters or "sub-themes" were discovered automatically by the LLM. Each subject's response was mapped exclusively to one primary cluster and could furthermore be associated with one or more secondary clusters.

LLM Response Validation for Clustering: The LLM returned two lists in its response: one of the cluster labels and their definitions, and the other of the cluster-labels (single or two-level clustering) assigned to each subject. The LLM was called once for each of 40 questions, and these responses were parsed using the Python Pandas library. A LLM response is marked invalid if it was ill-formatted (not in tab-separated-values format) or incomplete. The LLM was called again on invalid responses till the LLM returned a valid response. We found that almost 20 questions would have invalid responses in the first attempt, and in total, we were making almost 80% additional calls to get valid responses for all questions. We suspect that the rate of invalid responses in this step is higher than in the previous step due to the added complexity of the task i.e. the response needs to first produce a valid clustering-schema, and then additionally assign each of 93 subjects to the clusters according to the clustering schema.

Recluster: We repeated the above clustering step four additional times using a prompt similar to the previous clustering prompt (detailed in Appendix G). In this reclustering step, we used the same cluster definitions as were used in the previous steps, that is, a mix of expert-defined and LLM-generated (but expert-reviewed) cluster definitions. As in the original clustering, any invalid LLM responses were automatically detected and re-processed until a valid response was obtained. For the final cluster assignments used in downstream analysis, we applied a majority vote rule based on the 5 clustering repetitions. That is, each subject was assigned to the cluster they most commonly belonged to across the trials. This approach helps make the cluster assignments robust to the occasional variability in the LLM outputs. In a few cases ($< 1\%$ of all subject-question pairs), this process failed to find any cluster assignments that passed the majority-vote.

Human evaluation of LLM responses

Our study employed human evaluation to verify the alignment between RACER-generated clusters and human interpretation, utilizing two independent evaluators who analyzed the responses of 20 randomly selected subjects from a pool of 93. Each evaluator individually reviewed the raw interview transcript files for the selected 20 subjects and used the same cluster definitions as RACER to assign subjects to clusters. Human evaluators spent approximately 30 minutes per subject on average for a comprehensive review and categorization of the responses. This time investment reflects the thoroughness and attention to detail applied by the evaluators in their analysis, and also highlights the limits of this process to scale to large study populations. To validate the semantic clustering results produced by the LLM, each human evaluator compared their assigned scores with those generated by the LLM. An inter-rater comparison was also conducted, involving a detailed examination of the scores and evaluations independently made by both human evaluators (E1 and E2) for the same set of subjects. Concordance scores of 1 were assigned to clusters that precisely matched or were sub- or super-sets of each other, while discrepancies received a concordance score of 0. The overall concordance ratio represented the proportion of clusters aligning between the evaluators.

Additionally, the evaluators' findings were juxtaposed with RACER's cluster assignments to gauge both inter-evaluator consistency and the degree of correspondence with the LLM's outcomes. We also compared the use of Cohen's kappa coefficient with our concordance score and found them to be similar. Due to the nature of the comparison across questions which varied in the number of possible clusters as well as probability of different cluster assignment across questions, the concordance scores were used as they better described the intended comparisons. Instances where RACER did not produce any robust cluster assignments were categorized as 'mismatch' during the evaluation process.

B Insights using RACER on healthcare worker experience during COVID-19

Here we summarize the insights gleaned from analyzing SSIs with 93 subjects using RACER.

COVID-19 exposure, response, work impact and work changes:

The vast majority of practicing healthcare professionals reported having professional contact with COVID-19 patients in the past two months. Most subjects expressed safety concerns for themselves and loved ones, especially regarding viral exposure risks. Common protective measures adopted included heightened hygiene practices, using personal protective equipment, limiting travel and social interactions, and modifying routines at work and home to minimize transmission risks. Over half of the subjects reported physical tolls from the crisis, frequently citing exhaustion, disturbed sleep, and dietary changes (Figure 4).

Most subjects felt personally prepared to handle the pandemic, attributing this largely to their medical knowledge, experience, and ability to adapt. Assessment of institutional preparedness was more varied, with around 60% expressing their hospital/unit was prepared, but around 25% felt improvements were still needed.

Working hours markedly increased for most subjects during the pandemic, with over 80% reporting working more than 40 hours per week compared to pre-COVID times. For many, this resulted from escalations in patient load and administrative duties. Approaches to patient management also evolved, with the vast majority of practicing healthcare professionals stating their methods differed from usual practices. This included increased reliance on technology, more precautions with patients, and adjustments to treatments due to COVID-19. Most still felt capable of handling the situation professionally, though some desired more protections and support systems.

Among students and trainees, the majority believed they adhered closely to the Hippocratic oath during the pandemic. Their views on their educational institution's policies regarding medical students' roles during that time were divided, with half in agreement and others expressing mixed or negative sentiments, reflecting a spectrum of perspectives on the adequacy and effectiveness of institutional responses to the crisis.

Emotional and psychological impact, and support and coping strategies

The COVID-19 crisis negatively affected the emotional state of most subjects, with many reporting feelings of anxiety, stress, sadness, or anger. However, around 25% indicated a mix of both positive emotions like gratitude as well as negative feelings. Despite those challenges, the overwhelming majority felt supported by peers and family, suggesting strong social networks within and outside the workplace. Family dynamics had been affected for some, with around a quarter reporting increased family problems during the pandemic. This data underscored the profound emotional and psychological effects of the crisis on healthcare professionals, juxtaposed with the resilience and support systems that helped them navigate these challenges.

In regards to burnout, over 60% of subjects assessed their pre-pandemic burnout as low or mild. When asked about current burnout, around 40% still reported mild or no burnout, but the percentage reporting severe burnout rose from around 15% pre-pandemic to 20% during the crisis. If feeling burned out, nearly 90% stated they would seek help, with most mentioning professional resources like counseling. Over 60% also reported they would seek professional help if feeling mentally overwhelmed, with therapists and workplace programs being commonly cited options. However, around 45% still anticipated obstacles in getting help, including logistical barriers and stigma concerns (Figure 5).

Future considerations and professional outlook

When asked about near-term impacts, over 50% expressed concerns about anticipated difficulties, health risks, economic instability, and significant lifestyle changes. However, around 15% hoped for new opportunities and growth resulting from the crisis. Looking 5 years ahead, around 20% expected advancements in healthcare practices and systems due to learned lessons. Though nearly 10% feared lingering personal and professional impact. Among non-students considering job changes, around 15% expressed an immediate willingness to switch fields while around 18% would change contingent on worsening conditions.

Regarding effects on career plans, 35% of students reported the crisis has impacted their spe-

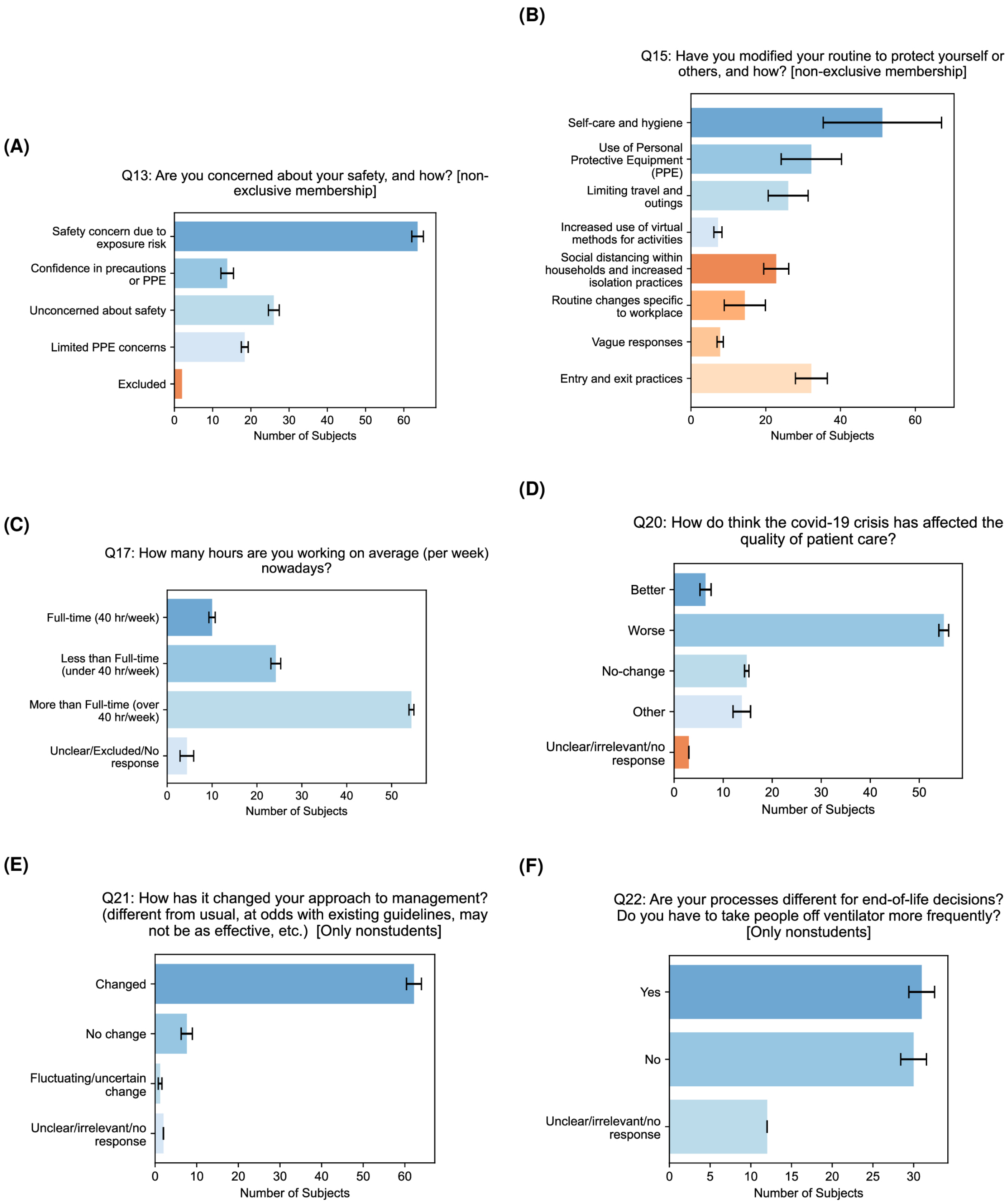

Figure 4: **Aggregated interview responses to selected questions about safety concerns arising from COVID-19 exposure, work impact, and medical management decisions.** Error bars reflect cluster-assignment variability arising from re-clustering step in RACER. Bar plot labels are primary clusters.

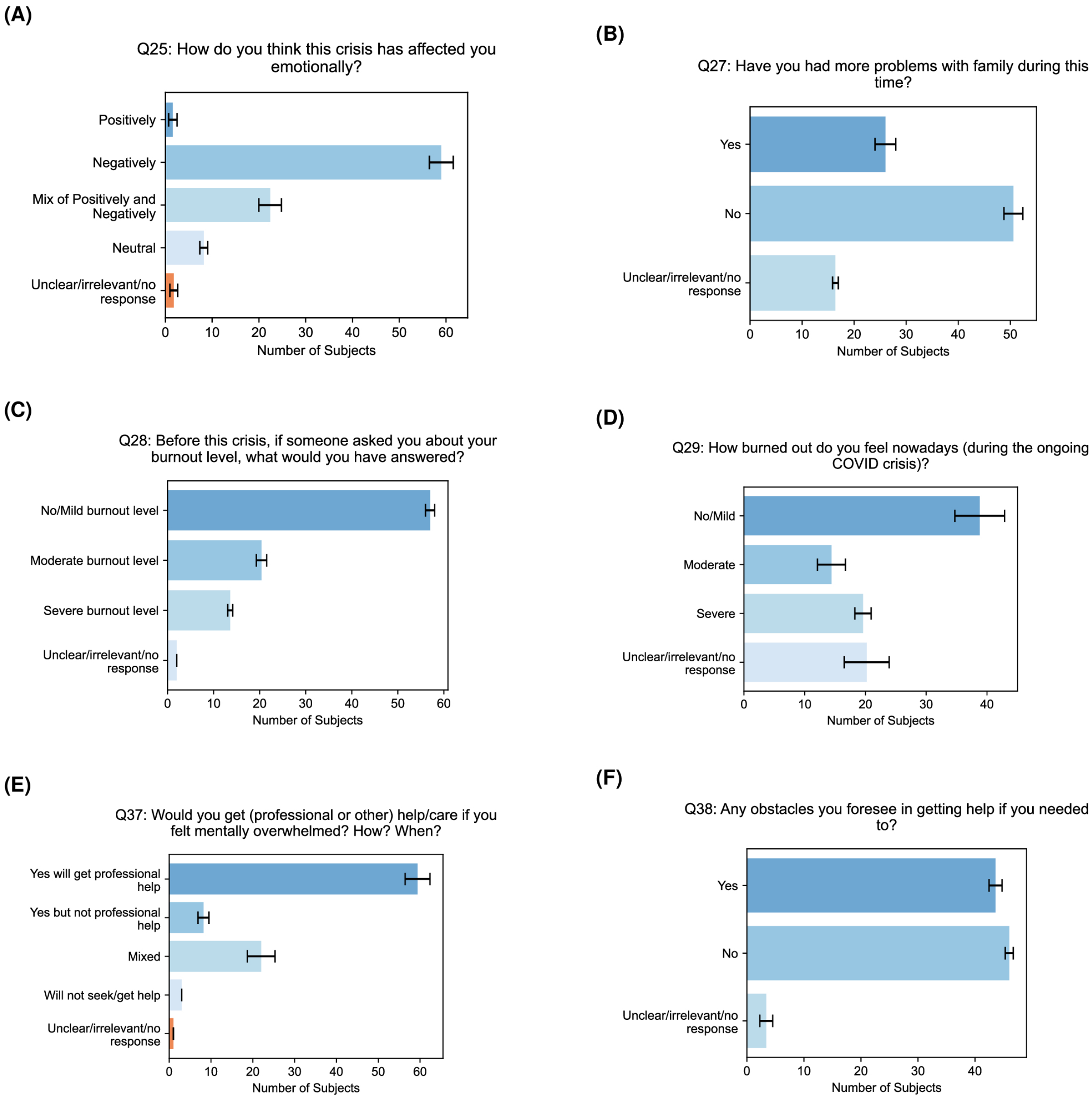

Figure 5: **Aggregated interview responses to selected questions about emotional and psychological impact, and support and coping strategies.** Error bars reflect cluster-assignment variability arising from re-clustering step in RACER. Bar plot labels are primary clusters.

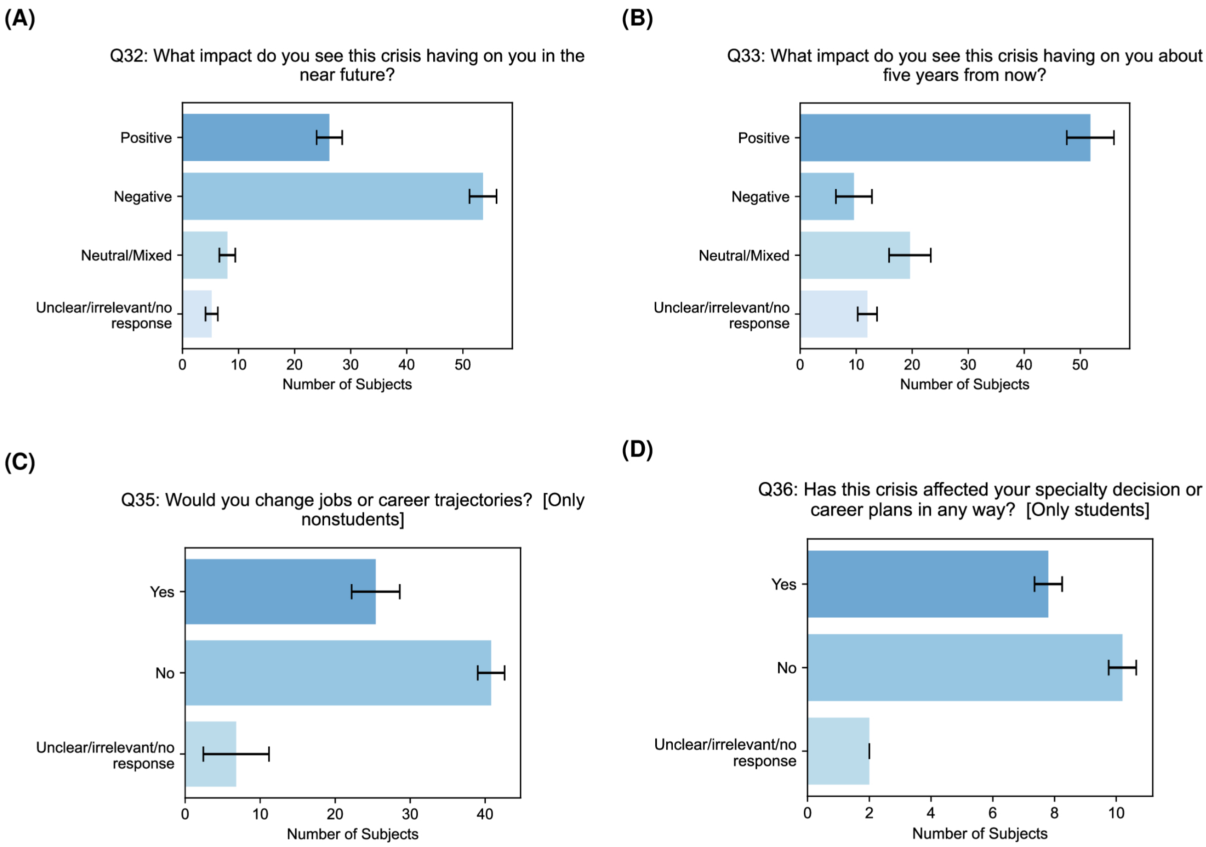

Figure 6: **Aggregated interview responses to selected questions about future considerations and professional outlook, as it relates to working in healthcare during or after the pandemic.** Error bars reflect cluster-assignment variability arising from re-clustering step in RACER. Bar plot labels are primary clusters.

cialty choices or work preferences. Specifically, around 20% described reconsidering their specialty choice due to the pandemic. Another 15% mentioned shifting their preferences regarding research involvement, practice locations, and other factors. However, 50% of students stated the crisis has not affected their professional plans or specialty decisions. Over 50% of students explicitly stated adherence to their Hippocratic oath obligations, while 10% conveyed adherence through descriptions of their clinical actions and interventions. Of students agreeing with their school's pandemic policies, 40% expressed unqualified agreement and 10% provided positive justifications. However, around 15% agreed tentatively due to concerns over student safety and curriculum changes (Figure 6).

C Interview questions and associated expert-guided primary clusters

Table 2: Expert provided primary clusters for questions.
Q1-Q13 and Q18 are *factual*, remaining are *subjective*. Q14-41 underwent human evaluation.

Q#	Question	Top-Level Cluster Guidance
1	How old are you?	(1) Young Adults (22 to 33), (2) Middle-aged Adults (34 to 45), (3) Older Adults (46 to 60), (4) Seniors (61 and above), and (5) Unclear/irrelevant/no response
2	Where do you live?	(1) Houston, TX, (2) San Antonio, TX, (3) TX (Other), (4) Florida, (5) Mid-West US, (6) US (Other) and (7) Unclear/Excluded/No response
3	What is your marital status?	(1) Not currently married, (2) Married currently, and (3) Unclear/Excluded/No response
15	Are you concerned about safety of loved ones, and how?	(1) Yes, (2) No, and (3) Unclear/irrelevant/no response
17	Has this crisis taken a toll on you physically in any way?	(1) Yes, (2) No, and (3) Unclear/irrelevant/no response
18	How many hours are you working on average (per week) nowadays?	(1) Full-time, (2) Less than Full-time, (3) More than Full-time, and (4) Unclear/Excluded/No response
19	How has your working schedule and logistics changed?	(1) Increased hours, (2) Decreased hours, (3) No change, (4) Other, and (5) Unclear/irrelevant/no response
20	How do your working hours compare to pre-covid-19 crisis?	(1) Increased hours, (2) Decreased hours, (3) No change, (4) Other, and (5) Unclear/irrelevant/no response
21	How do think the covid-19 crisis has affected the quality of patient care?	(1) Better, (2) Worse, (3) No-change, (4) Other and (5) Unclear/irrelevant/no response
22	How has it changed your approach to management?	(1) Changed, (2) No change, (3) Fluctuating/uncertain change, and (4) Unclear/irrelevant/no response
23	Are your processes different for end-of-life decisions? Do you have to take people off ventilator more frequently?	(1) Yes, (2) No, and (3) Unclear/irrelevant/no response
24	How prepared do you feel for the COVID-19 pandemic on a personal level?	(1) Prepared, (2) Unprepared, and (3) Unclear/irrelevant/no response
25	How prepared do you feel the unit/hospital is for the COVID-19 pandemic?	(1) Prepared, (2) Unprepared, and (3) Unclear/irrelevant/no response
26	How do you think this crisis has affected you emotionally?	(1) Positively (e.g. excitement), (2) Negatively, (3) Mix of Positively and Negatively, (4) Neutral, and (5) Unclear/irrelevant/no response
27	Do you feel supported by peers and/or family during this time?	(1) Yes, (2) No, (3) Mixed, (4) Fluctuating over time and (5) Unclear/irrelevant/no response
28	Have you had more problems with family during this time?	(1) Yes, (2) No, and (3) Unclear/irrelevant/no response
29	Before this crisis, if someone asked you about your burnout level, what would you have answered?	(1) No/Mild (e.g. 1, 2 or 3 out of 10), (2) Moderate (e.g. 4, 5 or 6 out of 10), (3) Severe (e.g. 7, 8, 9 or 10 out of 10), and (4) Unclear/irrelevant/no response

Continued on next page

Table 2 continued from previous page

Q#	Question	Top-Level Cluster Guidance
30	How burned out do you feel nowadays (during the ongoing COVID crisis)?	(1) No/Mild (e.g. 1, 2 or 3 out of 10), (2) Moderate (e.g. 4, 5 or 6 out of 10), (3) Severe (e.g. 7, 8, 9 or 10 out of 10), and (4) Unclear/irrelevant/no response
31	How do you feel about working from home OR at the frontlines?	(1) Positively (e.g. excitement), (2) Negatively, (3) Neutral/Mixed and (4) Unclear/irrelevant/no response
32	Do you feel you should be able to handle this as a healthcare professional?	(1) Yes, (2) No, (3) Mixed, and (4) Unclear/irrelevant/no response
33	What impact do you see this crisis having on you in the near future?	(1) Positive, (2) Negative, (3) Neutral/Mixed and (4) Unclear/irrelevant/no response
34	What impact do you see this crisis having on you about five years from now?	(1) Positive, (2) Negative, (3) Neutral/Mixed and (4) Unclear/irrelevant/no response
35	Would you seek help if you felt burned out? How?	(1) Yes, (2) No, and (3) Unclear/irrelevant/no response
36	Would you change jobs or career trajectories?	(1) Yes, (2) No, and (3) Unclear/irrelevant/no response
37	Has this crisis affected your specialty decision or career plans in any way?	(1) Yes, (2) No, and (3) Unclear/irrelevant/no response
38	Would you get (professional or other) help/care if you felt mentally overwhelmed? How? When?	(1) Yes will get professional help, (2) Yes but not professional help, (3) Mixed, (4) Will not seek/get help and (5) Unclear/irrelevant/no response
39	Any obstacles you foresee in getting help if you needed to?	(1) Yes, (2) No, and (3) Unclear/irrelevant/no response
40	If student or trainee, how closely do you feel that you are adhering to the Hippocratic oath during this time?	(1) Adhering Closely, (2) Not adhering closely OR Adhering conditionally, and (3) Unclear/irrelevant/no response
41	If student or trainee, do you agree with your school's policies regarding medical students' roles at this time?	(1) Yes, (2) No, (3) Mixed/Conditionally, and (4) Unclear/irrelevant/no response

D Interview questions with LLM-discovered primary clusters

Table 3: LLM-discovered (but expert-reviewed) Primary Clusters for remaining questions. Q1-Q13 and Q18 are *factual*, remaining are *subjective*. Q14-41 underwent human evaluation.

Q#	Question	LLM-discovered Primary Clusters
4	Do you have kids?	(1) Parents, (2) Non-parents, (3) Excluded
5	If you do have kids, provide details	[Non-exclusive membership] (1) No Information, (2) Single Child, (3) Two Children, (4) Three Children, (5) Four or more Children, (6) Child Age Provided, (7) Child Age Not Provided, (8) Children Living at Home, (9) Children No Longer Living at Home
6	Are you a caretaker otherwise? (if not own kids, e.g., elderly parents, adopted family member, etc.)	(1) Caretakers of Family Members, (2) Caretakers of Animals, (3) Partial Caretakers, (4) Financially Supportive, (5) No Caretaking Responsibilities, (6) Excluded

Continued on next page

Table 3 continued from previous page

Q#	Question	LLM-discovered Primary Clusters
7	What type of healthcare professional or student/trainee are you?	[Non-exclusive membership] (1) Physicians, (2) Medical Students, (3) Nurses, (4) Healthcare Professionals, (5) Residents, (6) Excluded
8	If student or trainee, what year are you in?	(1) First Year, (2) Second Year, (3) Third Year, (4) Fourth Year, (5) Unclear Training Year, (6) Excluded
9	What institution did you complete your (or are currently) training at?	[Non-exclusive membership] (1) Baylor College of Medicine, (2) University of Texas, (3) Texas Institutions, (4) Multiple Institutions, (5) Out of US Training, (6) Unspecified or Missing Information
10	If you are a physician, did you train in the US at any point?	(1) Trained in US, (2) Did not train in US, (3) No clear response
11	What is your specialty (if student, what specialty are you thinking of)?	[Non-exclusive membership] (1) Cardiology/Respiratory, (2) Neurology/Neurocritical Care, (3) Pediatrics, (4) Head and neck surgery/Related Surgery, (5) Fertility, (6) Psychiatry, (7) Emergency Medicine, (8) Pulmonary Critical Care, (9) Oncology, (10) OBGYN, (11) Infectious Diseases, (12) Anesthesiology and Critical Care, (13) Surgery/ER, (14) Internal Medicine, (15) Pathology, (16) Excluded
12	How long have you been practicing?	(1) Years under 15, (2) Years 15-30, (3) Years over 30, (4) Excluded
13	Over the past two months, have you practiced clinically in areas where you could be in touch with patients who have COVID-19?	(1) COVID-19 Patient Contact, (2) No COVID-19 Patient Contact
14	Are you concerned about your safety, and how?	[Non-exclusive membership] (1) Safety concern due to exposure risk, (2) Confidence in precautions or PPE, (3) Unconcerned about safety, (4) Limited PPE concerns
16	Have you modified your routine to protect yourself or others, and how?	[Non-exclusive membership] (1) Self-care and hygiene, (2) Use of Personal Protective Equipment (PPE), (3) Limiting travel and outings, (4) Increased use of virtual methods for activities, (5) Social distancing within households and increased isolation practices, (6) Routine changes specific to workplace, (7) Vague responses, (8) Entry and exit practices

E Prompt 1: Retrieving relevant responses from interview transcripts

```
Here is a template (tab-separated-values) of an interview (conducted
    ↪ in 2020) between an interviewer and a healthcare professional
    ↪ or medical student.
Populate the 'answer' column of the template below using the
    ↪ interview transcript appended after the template.
Be sure to note any positive, negative or neutral emotions expressed
    ↪ by the interviewee in the answer.
If a template question was not asked in the appended transcript (or
    ↪ is not applicable), the answer should be "NA".
For the last 'evidence' column, provide evidence, by quoting verbatim
    ↪  (except for newlines) the parts of the transcript that were
    ↪ most relevant to answering the question.

question_number question        answer  evidence
1       How old are you?        [numeric]
2       Where do you live?      [city, state, country]
3       What is your marital status?    [single/married/divorced/
    ↪ widowed/etc]
4       Do you have kids?       [yes/no]
5       If you do have kids, provide details    [details]
6       Are you a caretaker otherwise? (if not own kids, eg elderly
    ↪ parents, adopted family member, etc)        [yes/no; details]
7       What type of healthcare professional or student/trainee are
    ↪ you?        [details]
8       If student or trainee, what year are you in?    [year of
    ↪ program]
9       What institution did you complete your (or are currently)
    ↪ training at?  [name and location of institution]
10      If you are a physician, did you train in the US at any point?
    ↪   [yes/no]
11      What is your specialty (if student, what specialty are you
    ↪ thinking of)?        [details]
12      How long have you been practicing?      [in years, or NA for
    ↪ student]
13      Over the past two months, have you practiced clinically in
    ↪ areas where you could be in touch with patients who have covid
    ↪ -19?   [yes/no]
14      Are you concerned about your safety, and how?   [yes/no;
    ↪ details]
15      Are you concerned about safety of loved ones, and how?  [yes/
    ↪ no; details]
16      Have you modified your routine to protect yourself or others,
    ↪  and how?  [yes/no; details]
17      Has this crisis taken a toll on you physically in any way?
    ↪       [yes/no; details]
18      How many hours are you working on average (per week) nowadays
    ↪ ?  [numeric]
19      How has your working schedule and logistics changed?    [
    ↪ details]
20      How do your working hours compare to pre-covid-19 crisis?
```

```
        [details]
21    How do think the covid-19 crisis has affected the quality of patient care?     [details]
22    How has it changed your approach to management? (different from usual, at odds with existing guidelines, may not be as effective, etc.) [details]
23    Are your processes different for end-of-life decisions? Do you have to take people off ventilator more frequently?     [details]
24    How prepared do you feel for the COVID-19 pandemic on a personal level? [details]
25    How prepared do you feel the unit/hospital is for the COVID-19 pandemic?        [details]
26    How do you think this crisis has affected you emotionally?      [note emotions recognized from interviewee;details]
27    Do you feel supported by peers and/or family during this time?  [details]
28    Have you had more problems with family during this time?        [details]
29    Before this crisis, if someone asked you about your burnout level, what would you have answered?        [score (e.g. 6 out of 10) and/or details]
30    How burned out do you feel nowadays (during the ongoing COVID crisis)?  [score (e.g. 6 out of 10) and/or details]
31    How do you feel about working from home OR at the frontlines?   [Home/Frontlines/Other; details]
32    Do you feel you should be able to handle this as a healthcare professional?     [yes/no; details]
33    What impact do you see this crisis having on you in the near future?    [details]
34    What impact do you see this crisis having on you about five years from now?     [details]
35    Would you seek help if you felt burned out? How?        [yes/no; details]
36    Would you change jobs or career trajectories?   [yes/no; details]
37    Has this crisis affected your specialty decision or career plans in any way?    [yes/no; details]
38    Would you get (professional or other) help/care if you felt mentally overwhelmed? How? When?    [yes/no; details]
39    Any obstacles you foresee in getting help if you needed to?     [yes/no; details]
40    If student or trainee, how closely do you feel that you are adhering to the Hippocratic oath during this time?  [closely/not-closely; details]
41    If student or trainee, do you agree with your school's policies regarding medical students' roles at this time? [yes/no; details]
```

TRANSCRIPT:

[Interview Transcript Appended]

F Prompt 2: Template for semantic Clustering of responses aggregated across all subjects

Out of 41 questions in our template in E, 29 questions had expert-provided templates that defined the primary clusters but left secondary-cluster definitions to the LLM. Two questions (Q14, Q16) used LLM-discovered (but expert-reviewed) single-level clustering with non-exclusive membership. The following Python code shows the template used for generating the prompt associated with each question (note the use of zero-indexing):

```
TEMPLATE = """Cluster the responses in the table below at two levels.
Top level clusters must be {clusters}.
Top level clusters have mutually-exclusive cluster membership.
For the next level, cluster the responses from subjects belonging to
   ↪ each top-level cluster highlighting the common theme per
   ↪ cluster.
Subjects can belong to multiple clusters at this level.

Your response should be in tab-separated-values format, with the
   ↪ following columns:
subject_id  top_level_cluster_id    secondary_cluster_ids

Example output line:
C-002   C1  "C1.1,C1.2,C1.4"

Start your response by defining each top and secondary cluster in tab
   ↪ -separated-values format, with columns:
cluster_id  cluster_name    cluster_description

Note that some subject_ids may not be present in the prompt, and so
   ↪ should also not be present in your response.
Provide both the (tab-separated) cluster-definitions table and the (
   ↪ tab-separated) cluster-assignments table in your response.
\n"""

prompts = {
    "default": """Cluster the responses in the table below
       ↪ highlighting the common theme per cluster.
Group subjects that provide unclear, irrelevant, or no responses into
   ↪  a separate "excluded" cluster.
Subjects can belong to multiple clusters. Your response should be in
   ↪ tab-separated-values format,
with the following columns: subject_id, cluster_ids

Example output line:
subject_id  cluster_ids
C-002   "C2,C3"

Start your response by defining each cluster in tab-separated-values
   ↪ format, with columns:
cluster_id, cluster_name, cluster_description

Note that some subject_ids may not be present in the prompt, and so
   ↪ should also not be present in your response.
```

```
Provide both the (tab-separated) cluster-definitions table and the (
    ↪ tab-separated) cluster-assignments table in your response.
\n""",
    0: TEMPLATE.format(
        clusters="(1) Young Adults (22 to 33), (2) Middle-aged Adults
            ↪  (34 to 45), (3) Older Adults (46 to 60), (4) Seniors
            ↪ (61 and above), and (5) Unclear/irrelevant/no response"
    ),
    1: TEMPLATE.format(
        clusters="(1) Houston, Texas, (2) San Antonio, Texas, (3)
            ↪ Texas (Other), (4) Florida, (5) Mid-West US, (6) US (
            ↪ Other) and (7) Unclear/Excluded/No response"
    ),
    2: TEMPLATE.format(
        clusters="(1) Not currently married, (2) Married currently,
            ↪ and (3) Unclear/Excluded/No response"
    ),
    14: TEMPLATE.format(
        clusters="(1) Yes, (2) No, and (3) Unclear/irrelevant/no
            ↪ response"
    ),
    16: TEMPLATE.format(
        clusters="(1) Yes, (2) No, and (3) Unclear/irrelevant/no
            ↪ response"
    ),
    # 17: Numeric: How many hours are you working on average (per
        ↪ week)?
    17: TEMPLATE.format(
        clusters="(1) Full-time, (2) Less than Full-time, (3) More
            ↪ than Full-time, and (4) Unclear/Excluded/No response"
    ),
    18: TEMPLATE.format(
        clusters="(1) Increased hours, (2) Decreased hours, (3) No
            ↪ change, (4) Other, and (5) Unclear/irrelevant/no
            ↪ response"
    ),
    # 19: How does this compare to pre-covid-19 crisis?
    19: TEMPLATE.format(
        clusters="(1) Increased hours, (2) Decreased hours, (3) No
            ↪ change, (4) Other, and (5) Unclear/irrelevant/no
            ↪ response"
    ),
    20: TEMPLATE.format(
        clusters="(1) Better, (2) Worse, (3) No-change, (4) Other and
            ↪  (5) Unclear/irrelevant/no response"
    ),
    21: TEMPLATE.format(
        clusters="(1) Changed, (2) No change, (3) Fluctuating/
            ↪ uncertain change, and (4) Unclear/irrelevant/no
            ↪ response"
    ),
    22: TEMPLATE.format(
```

```
    clusters="(1) Yes, (2) No, and (3) Unclear/irrelevant/no
       ↪ response"
),
23: TEMPLATE.format(
    clusters="(1) Prepared, (2) Unprepared, and (3) Unclear/
       ↪ irrelevant/no response"
),
24: TEMPLATE.format(
    clusters="(1) Prepared, (2) Unprepared, and (3) Unclear/
       ↪ irrelevant/no response"
),
25: TEMPLATE.format(
    clusters="(1) Positively (e.g. excitement), (2) Negatively,
       ↪ (3) Mix of Positively and Negatively, (4) Neutral, and
       ↪ (5) Unclear/irrelevant/no response"
),
26: TEMPLATE.format(
    clusters="(1) Yes, (2) No, (3) Mixed, (4) Fluctuating over
       ↪ time and (5) Unclear/irrelevant/no response"
),
27: TEMPLATE.format(
    clusters="(1) Yes, (2) No, and (3) Unclear/irrelevant/no
       ↪ response"
),
28: TEMPLATE.format(
    clusters="(1) No/Mild (e.g. 1, 2 or 3 out of 10), (2)
       ↪ Moderate (e.g. 4, 5 or 6 out of 10), (3) Severe (e.g.
       ↪ 7, 8, 9 or 10 out of 10), and (4) Unclear/irrelevant/no
       ↪  response"
),
29: TEMPLATE.format(
    clusters="(1) No/Mild (e.g. 1, 2 or 3 out of 10), (2)
       ↪ Moderate (e.g. 4, 5 or 6 out of 10), (3) Severe (e.g.
       ↪ 7, 8, 9 or 10 out of 10), and (4) Unclear/irrelevant/no
       ↪  response"
),
30: TEMPLATE.format(
    clusters="(1) Positively (e.g. excitement), (2) Negatively,
       ↪ (3) Neutral/Mixed and (4) Unclear/irrelevant/no
       ↪ response"
),
31: TEMPLATE.format(
    clusters="(1) Yes, (2) No, (3) Mixed, and (4) Unclear/
       ↪ irrelevant/no response"
),
32: TEMPLATE.format(
    clusters="(1) Positive, (2) Negative, (3) Neutral/Mixed and
       ↪ (4) Unclear/irrelevant/no response"
),
33: TEMPLATE.format(
    clusters="(1) Positive, (2) Negative, (3) Neutral/Mixed and
       ↪ (4) Unclear/irrelevant/no response"
```

```
    ),
    34: TEMPLATE.format(
        clusters="(1) Yes, (2) No, and (3) Unclear/irrelevant/no
          ↪ response"
    ),
    35: TEMPLATE.format(
        clusters="(1) Yes, (2) No, and (3) Unclear/irrelevant/no
          ↪ response"
    ),
    36: TEMPLATE.format(
        clusters="(1) Yes, (2) No, and (3) Unclear/irrelevant/no
          ↪ response"
    ),
    37: TEMPLATE.format(
        clusters="(1) Yes will get professional help, (1) Yes but not
          ↪  professional help, (3) Mixed, (4) Will not seek/get
          ↪ help and (5) Unclear/irrelevant/no response"
    ),
    38: TEMPLATE.format(
        clusters="(1) Yes, (2) No, and (3) Unclear/irrelevant/no
          ↪ response"
    ),
    39: TEMPLATE.format(
        clusters="(1) Adhering Closely, (2) Not adhering closely OR
          ↪ Adhering conditionally, and (3) Unclear/irrelevant/no
          ↪ response"
    ),
    40: TEMPLATE.format(
        clusters="(1) Yes, (2) No, (3) Mixed/Conditionally, and (3)
          ↪ Unclear/irrelevant/no response"
    ),
}
```

G Prompt 3: Re-Clustering using previously defined clusters

```
Cluster the responses in the table below highlighting the common
    ↪ theme per cluster.
Group subjects that provide unclear, irrelevant, or no responses into
    ↪  a separate "excluded" cluster.
Subjects can belong to multiple clusters. Your response should be in
    ↪ tab-separated-values format,
with the following columns: subject_id, cluster_ids

Example output line:
subject_id  cluster_ids
C-002    "C2,C3"

Note that some subject_ids may not be present in the prompt, and so
    ↪ should also not be present in your response.
Provide both the (tab-separated) cluster-definitions table and the (
    ↪ tab-separated) cluster-assignments table in your response.

subject_id      Are you a caretaker otherwise? (if not own kids, eg
    ↪ elderly parents, adopted family member, etc)
C001     No
C002     No
C003     No
C004     No
C005     No

...

C086     Yes, looks after his mother-in-law's finances
C087     No
C090     Yes; Partial caretaker for parents
C099     No
C100     No
C101     No
C102     No

Use the following cluster definitions (Do not repeat this in output):
cluster_id      cluster_name     cluster_description
C1       Caretakers of Family Members    Subjects who responded that
    ↪ they take care of relatives (elderly parents, children,
    ↪ siblings or others).
C2       Caretakers of Animals    Subjects who take care of animals.
C3       Partial Caretakers       Subjects who participate in
    ↪ caretaking but not as primary caretakers.
C4       Financially Supportive   Subjects who provide financial
    ↪ support instead of physical caretaking.
C5       No Caretaking Responsibilities   Subjects who stated that they
    ↪  do not take care of anyone.
C6       Excluded         Responses that are unclear, irrelevant, or
    ↪ did not provide a response to the question.
```