

---

# Demo: Guide-RAG: Evidence-Driven Corpus Curation for Retrieval-Augmented Generation in Long COVID

---

**Philip DiGiacomo**

Department of Computer Science  
University of Texas at Austin

**Haoyang Wang**

School of Information  
University of Texas at Austin

**Jinrui Fang**

School of Information  
University of Texas at Austin

**Yan Leng**

McCombs School of Business  
University of Texas at Austin

**W Michael Brode**

Dell Medical School  
University of Texas at Austin

**Ying Ding**

School of Information  
University of Texas at Austin

## Abstract

As AI chatbots gain adoption in clinical medicine, developing effective frameworks for complex, emerging diseases presents significant challenges. We developed and evaluated six Retrieval-Augmented Generation (RAG) corpus configurations for Long COVID (LC) clinical question answering, ranging from expert-curated sources to large-scale literature databases. Our evaluation employed an LLM-as-a-judge framework across faithfulness, relevance, and comprehensiveness metrics using LongCOVID-CQ, a novel dataset of expert-generated clinical questions. Our RAG corpus configuration combining clinical guidelines with high-quality systematic reviews consistently outperformed both narrow single-guideline approaches and large-scale literature databases. Our findings suggest that for emerging diseases, retrieval grounded in curated secondary reviews provides an optimal balance between narrow consensus documents and unfiltered primary literature, supporting clinical decision-making while avoiding information overload and oversimplified guidance. We propose Guide-RAG, a chatbot system and accompanying evaluation framework that integrates both curated expert knowledge and comprehensive literature databases to effectively answer LC clinical questions.

## 1 Introduction

Artificial intelligence (AI) tools and large language models (LLMs) are rapidly being adopted in clinical medicine, with both patients and clinicians increasingly turning to chatbot platforms for on-demand medical information [1–7]. For clinicians, resources such as OpenEvidence are emerging to augment traditional references like Harrison’s textbook or UpToDate online reference, offering algorithmically synthesized summaries of clinical literature [8, 9]. This rapid uptake underscores both the promise and the risk of chatbot-based decision support: while these systems can provide interactive, timely guidance, they remain vulnerable to hallucination, citation errors, and representational biases [1, 3–5, 10]. Designing best-practice frameworks for medical chatbots has

therefore become an urgent challenge, especially for diseases that are complex, heterogeneous, or poorly understood [3, 5, 11].

Long COVID (LC) represents the challenge of treating complex and emerging illnesses. Affecting an estimated 7% of U.S. adults (approximately 18 million people), LC is a heterogeneous, multisystem condition characterized by more than 200 reported symptoms, and lacks a standardized diagnostic biomarker or evidence-based treatment [12–14]. Clinical guidance remains largely consensus-based and is often extrapolated from related syndromes such as myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) and postural orthostatic tachycardia syndrome (POTS) [15–17]. In such high-uncertainty settings without clear ground truth, designing chatbots for clinical decision support is particularly challenging, as relying on PubMed alone or a single guideline risks overwhelming clinicians or missing evolving evidence. Retrieval-Augmented Generation (RAG) systems offer a promising approach to address these limitations by grounding language model responses in external evidence. The selection of corpus documents is critical to the system’s capability to answer clinical questions. In the case of emerging diseases, current approaches span two extremes: manually curated, high-quality databases versus large-scale literature collections that provide comprehensive coverage but introduce quality and relevance challenges. We propose a balanced approach using a novel combination of a LC consensus guideline with high-quality systematic reviews.

We evaluated RAG performance across three dimensions: faithfulness (ensuring responses are grounded in source documents), relevance (maintaining focus on the clinical question), and comprehensiveness (providing thorough coverage of complex, multisystem presentations). Our evaluation employed an LLM-as-a-judge framework that went beyond traditional accuracy metrics to assess criteria directly relevant to clinical practice. Additionally, we developed LongCOVID-CQ, a specialized dataset of expert-generated clinical questions targeting diagnosis, management strategies, and mechanisms that practicing clinicians routinely encounter when caring for LC patients. We summarize Guide-RAG’s three key contributions:

1. **Expert corpus curation:** Targeted curation of a clinical guideline supplemented by three high-quality systematic reviews outperformed both large-scale literature databases and narrow single-guideline approaches for LC question answering using RAG.
2. **Evaluation metrics:** We adopted faithfulness, relevance, and comprehensiveness metrics specifically for LC clinical applications using an LLM-as-a-judge framework to capture criteria directly relevant to clinical decision-making and trust-building.
3. **LongCOVID-CQ:** We developed a specialized evaluation dataset of expert-generated long COVID clinical questions (LongCOVID-CQ) reflecting the practical information needs that providers routinely encounter in patient care.

## 2 Related Work

**RAG applications to emerging and unknown diseases.** Emerging diseases present a dual challenge for information systems: the scarcity of high-quality data [18] and the overload of rapidly published yet conflicting information [19]. Compact, domain-specific knowledge bases can improve RAG performance via factual recall and diagnostic reasoning [20, 21]. However, quickly evolving research may not be captured by the curated knowledge base. Using large but unfiltered sources such as PubMed presents distinct challenges [22]: irrelevant or low-quality articles can degrade model accuracy, and the “lost-in-the-middle” effect can prevent models from recalling information buried within extensive text. Clinical guidelines are useful tools in navigating the specific versus general knowledge tradeoff, by distilling clinical consensus out of broad research topics. Guidelines for well-studied medical subjects have been successively applied to RAG pipelines in contrast-media consultations and perioperative assessments [23, 24]. We found that combining guidelines with systematic reviews widened coverage without re-introducing noise. To our knowledge, no published study has evaluated a RAG framework that grounds answers jointly in clinical guidelines and systematic reviews for LC or other emerging and unknown diseases.

**Evaluation of RAG for clinical relevance.** Existing medical question-answering (QA) benchmarks predominantly employ multiple-choice formats [25–28]. Consequently, medical-specific RAG chatbots are evaluated primarily on accuracy metrics when selecting from predetermined options [29–31]. This narrow focus on accuracy fails to capture critical aspects of clinical utility, including decision support capabilities and adherence to clinical guidelines [32, 33]. Recent work has begun

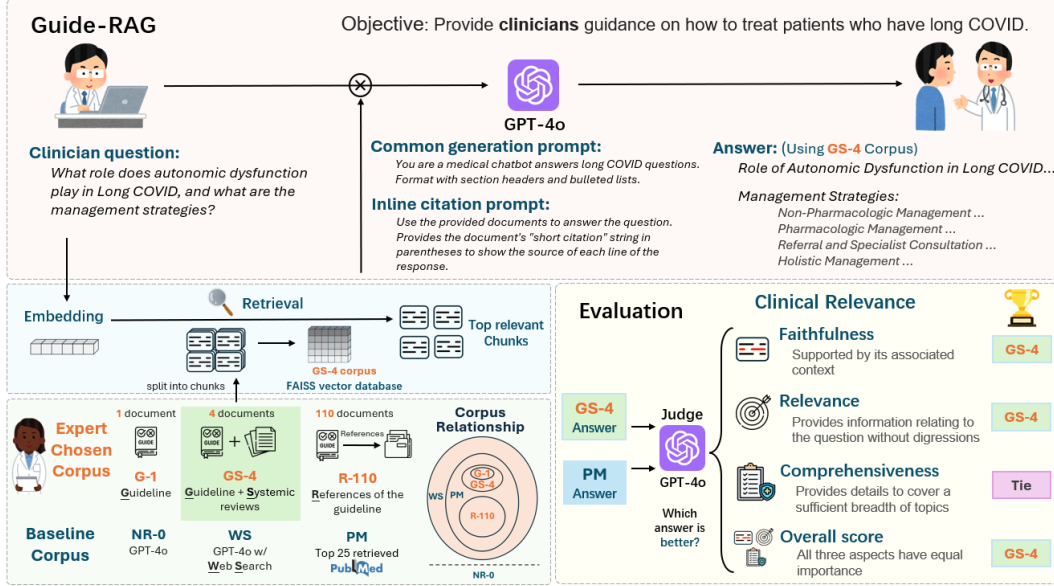


Figure 1: Workflow of Guide-RAG answering LC questions. The full prompt used for the answer generation and evaluation is presented in Appendix A.

exploring multi-dimensional evaluation of RAG systems in clinical settings using LLM-as-a-judge [34]. Wada et al. [23] evaluated response time, applicability, structure, safety, and professional communication, while Wu et al. [35] assessed pertinence, correctness, understandability, precision, and recall. However, the abundance of available metrics necessitates principled selection based on downstream application requirements. For LC clinical application, we identify three evaluation dimensions. First, faithfulness ensures clinician trust by requiring responses to be substantiated by source documents [36]. Second, given the 200+ reported symptoms of LC [12–14], chatbots must maintain relevance to presented symptoms [36]. Finally, comprehensiveness ensures holistic assessment across the vast symptom space [37]. We utilized an LLM-as-a-judge evaluation framework that assessed RAG systems across faithfulness, relevance, and comprehensiveness for LC queries. This approach went beyond accuracy-based metrics while targeting criteria essential for LC clinical practice.

### 3 Methods

#### 3.1 Corpus construction

We systematically evaluated six distinct corpus configurations within our Guide-RAG framework to assess the impact of document curation strategies on clinical question-answering performance (Figure 1). We performed our analysis across faithfulness, relevance, and comprehensiveness metrics, while including an overall metric that combines all three metrics equally. Our experimental design contrasts expert-curated knowledge bases with large-scale literature corpuses using both sparse and dense retrieval methods (detailed in Appendix C). This investigation addressed the clinical challenge where existing frameworks relying on vast, unfiltered literature repositories like PubMed may produce technically accurate but clinically misaligned outputs, particularly for emerging diseases with limited diagnostic clarity and therapeutic consensus.

**No retrieval (NR-0).** Standard GPT-4o without retrieval augmentation served as our control condition.

**Expert-curated corpuses.** Three configurations leveraged domain expertise from a physician-scientist specializing in LC care. **Guideline only (G-1).** This minimal configuration consisted solely of the American Academy of Physical Medicine and Rehabilitation (AAPM&R) "Multidisciplinary collaborative guidance on the assessment and treatment of patients with Long COVID: A compendium statement" [16]. This document represents the current consensus-driven guidance on LC in the United States, developed through a multi-year collaborative process by active researchers and clinicians.

**Guideline + systematic reviews (GS-4).** This configuration augmented G-1 with three high-quality systematic reviews selected for comprehensiveness and recency, providing the current synthesis of the evolving LC literature from leading medical journals. **References of the guideline (R-110).** This corpus comprised the 110 references cited within the AAPM&R guideline. We excluded one overlapping citation with GS-4 to ensure independent comparison [12] and two references were omitted due to the infeasibility of preprocessing as detailed in Appendix H. This configuration tested whether grounding the model in the evidence base underlying consensus guidance improved response quality relative to summary documents alone.

**Large-scale literature corpora.** We adopted vast bodies of available research as baselines for comparison. **PubMed corpus (PM).** This configuration accessed the comprehensive biomedical literature database maintained by the National Library of Medicine (NLM), encompassing MEDLINE (>39M citations), PubMed Central (full-text articles), and Bookshelf (biomedical books). The scale necessitated a hybrid sparse-dense retrieval approach (Appendix C). **Web search (WS).** This configuration employed OpenAI’s GPT-4o web search capabilities with explicit constraints to retrieve only peer-reviewed medical publications with source attribution.

### 3.2 Dataset: LongCOVID-CQ

A physician-scientist specializing in LC care developed 20 questions reflecting the key issues a practicing clinician is likely to encounter. Rather than a general summary, these queries target clinically relevant topics such as diagnosis, management strategies, and mechanisms to capture the kinds of questions providers routinely face in caring for patients with LC. These questions are listed in Appendix B.

### 3.3 Evaluation framework

**Evaluation metrics.** We employed three criteria, plus an overall metric, targeting responses that are useful for clinical application. **Faithfulness** measures whether generated content is directly supported by retrieved document chunks, enabling citation traceability essential for clinical trust and verification. **Relevance** quantifies the degree to which responses directly address the input question without extraneous content that burdens clinical workflow. **Comprehensiveness** evaluates coverage breadth, which is important given the >200 documented LC symptoms requiring multi-faceted responses. **Overall** performance aggregates faithfulness, relevance, and comprehensiveness with equal weighting to provide holistic quality assessment.

**LLM-as-a-Judge** We employed GPT-4o as an evaluation model for pairwise comparisons across all corpus configurations. For each input question, responses were generated from all six configurations and evaluated head-to-head across the three criteria mentioned, plus an overall metric. The evaluation prompt included explicit instructions for handling response length bias and positional effects, with randomized presentation order to mitigate systematic biases. All evaluation prompts included one-shot examples and explicit tie-handling instructions. Complete prompting details are provided in Appendix A. Win rates for each comparison were calculated by assigning a score of 100 to the winning response, 0 to the losing response, and 50 to both responses in case of a tie. These scores were then averaged over the questions in LongCOVID-CQ.

## 4 Results

**GS-4 achieved superior overall performance.** Our overall evaluation encompassing all three metrics demonstrated that GS-4 consistently outperformed competing approaches, achieving win rates of 57.5-65% in pairwise comparisons. This superior performance extended to two of the three individual metrics, with GS-4 ranking highest in both faithfulness and comprehensiveness evaluations. Notably, despite utilizing only 4 sources, GS-4 achieved greater comprehensiveness than both R-110 and PM (60% win rate for both comparisons), which leveraged significantly larger corpora. Win rate data is shown for all comparisons in Figure 2. Our heatmaps did not include NR-0 or WS for the faithfulness or overall comparisons because retrieved chunks of text were required to make those comparisons.

**PM exhibited strongest relevance performance.** While GS-4 dominated overall metrics, PM demonstrated slight superiority in relevance evaluation, achieving win rates of 50-52.5% against

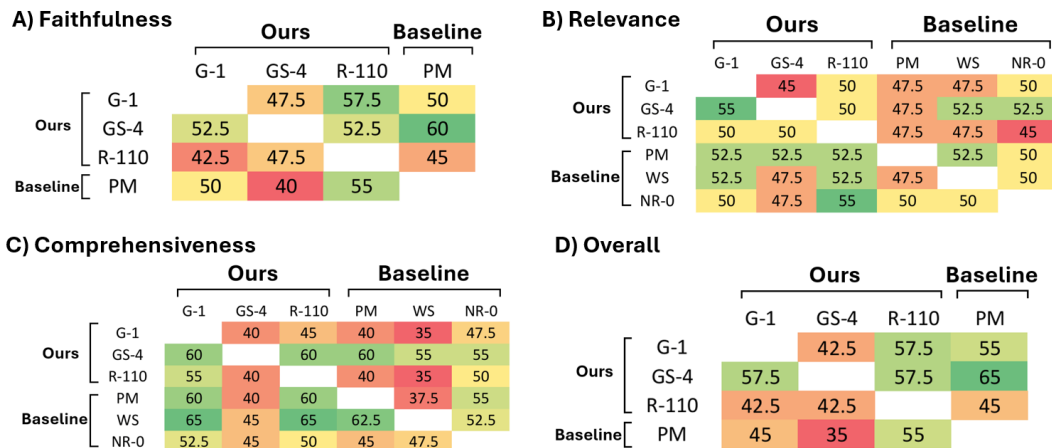


Figure 2: Win rates for (A) faithfulness, (B) relevance, (C) comprehensiveness, and (D) overall (each cell is the percentage of comparisons won by the row label over the column label)

other conditions. We hypothesize that PM’s access to expanded information enables this enhanced relevance performance.

**Guideline outperformed constituent references.** The comparison between G-1 and R-110 provided insight into the relative efficacy of synthesized guidelines versus their underlying references. G-1 achieved a 57.5% win rate over R-110 in faithfulness and overall evaluation. While R-110 maintained an advantage in comprehensiveness (55% win rate) and had comparable relevance performance (50% win rate), G-1’s overall superiority suggests that synthesized guidelines can provide enhanced faithfulness through better alignment with retrieved content, while being comprehensive and relevant enough to not degrade overall performance.

**Expert Evaluation.** Beyond the quantitative win-rate data, expert clinical review helps contextualize these results for real-world LC care. The heatmaps showed that corpora built from expert-curated sources consistently outperformed broader or unstructured collections, especially in domains like symptom management and mechanisms where overly specific or overconfident responses can be misleading. Figure 3 illustrates this with side-by-side outputs from different corpus conditions. Broader corpora (R-110 and PM) produced answers that appeared superficially plausible in their specificity but leaned heavily on single studies, and overemphasized speculative mechanisms. For example, R-110 attributed autonomic dysfunction solely to vagus nerve imbalance, and even recommended poorly researched interventions such as stellate ganglion block. The PM response went further, endorsing exercise training based on a review article that extrapolated treatment approaches from other diseases—advice that is explicitly cautioned against by LC experts, where standard exercise regimens may worsen symptoms. By contrast, the curated GS-4 corpus framed autonomic dysfunction as one part of a systemic illness and offered practical management guidance that acknowledged evidence gaps and suggested specialist involvement when appropriate. It also correctly noted that any physical activity must be carefully tailored to the patient. From a clinical perspective, this response is far more useful: it equips clinicians to counsel patients responsibly while avoiding misleading certainty. This example underscores how uncured data can generate outputs that appear authoritative yet are incomplete or misleading, and why expert evaluation remains essential alongside quantitative benchmarks when developing AI systems for emerging illnesses like LC. Full responses can be compared in Appendix D-F.

## 5 Discussion and Conclusion

These findings suggest a design principle for chatbots in emerging disease contexts: retrieval grounded in curated secondary reviews may provide a balance between narrow consensus documents and the unfiltered breadth of primary literature. The early COVID-19 pandemic illustrates the stakes, with more than 23,000 COVID-19 articles indexed in the first six months alone, many offering little new or meaningful information [38]. In such settings, indiscriminate retrieval risks producing

technically accurate but clinically misaligned outputs. Our evaluation provides preliminary evidence that systematic reviews and expert guidance can anchor chatbot responses in a way that supports faithfulness, relevance, and comprehensiveness: qualities that go beyond accuracy and are essential for decision support in high-uncertainty medical domains. The LongCOVID-CQ dataset addresses a critical gap in medical AI evaluation by providing clinically grounded questions that reflect real-world information needs. Unlike existing medical QA benchmarks that predominantly employ multiple-choice formats testing factual recall, our expert-generated questions target the diagnostic uncertainty, management complexity, and mechanistic understanding that practicing clinicians actually encounter when caring for LC patients.

## 6 Limitations

This study was limited by its focus on a single clinical domain (LC) and by the use of a single LLM model (GPT-4o) as the evaluation judge without validation against human or model-diverse raters. The small, expert-generated question set limited statistical robustness, and differences in retrieval strategies across corpora, such as dense retrieval for smaller datasets versus hybrid sparse-dense retrieval for PubMed, may have introduced bias in recall quality. The study also did not include systematic exploration of retrieval hyperparameters, including chunk size, embedding model selection, and reranking thresholds. These factors together constrain the generalizability and reproducibility of the findings.

Future work should include human expert ratings, multiple model evaluations, retrieval ablations, and will extend testing to additional high-complexity and high-uncertainty medical domains. By beginning with LC, a condition characterized by high-complexity and evolving evidence, this framework may provide generalizable principles for developing trustworthy AI systems across other complex diseases.

## References

- [1] Leyao Wang, Zhiyu Wan, Congning Ni, Qingyuan Song, Yang Li, Ellen Clayton, Bradley Malin, and Zhijun Yin. Applications and concerns of chatgpt and other conversational large language models in health care: systematic review. *Journal of Medical Internet Research*, 26:e22769, 2024.
- [2] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.
- [3] Bright Huo, Amy Boyle, Nana Marfo, Wimonchat Tangamornsuksan, Jeremy P. Steen, Tyler McKechnie, Yung Lee, Julio Mayol, Stavros A. Antoniou, Arun James Thirunavukarasu, Samantha Sanger, Kumail Ramji, and Gordon Guyatt. Large language models for chatbot health advice studies: a systematic review. *JAMA Network Open*, 8(2):e2457879–e2457879, 2025.
- [4] Rachel S. Goodman, J. Randall Patrinely, Cosby A. Stone, Eli Zimmerman, Rebecca R. Donald, Sam S. Chang, Sean T. Berkowitz, Avni P. Finn, Eiman Jahangir, Elizabeth A. Scoville, Thomas S. Reese, Debra L. Friedman, Julie A. Bastarache, Yuri F. van der Heijden, Justin J. Wright, Fei Ye, Nicole Carter, Meredith R. Alexander, Justin H. Choe, Carrie A. Chastain, John A. Zic, Sara N. Horst, Ilse Turker, Rajat Agarwal, Emily Osmundson, Kamran Idrees, Celia M. Kiernan, Chandrasekhar Padmanabhan, Christina E. Bailey, Catherine E. Schlegel, Lola B. Chambless, Michael K. Gibson, Travis J. Osterman, Lee E. Wheless, and Douglas B. Johnson. Accuracy and reliability of chatbot responses to physician questions. *JAMA network open*, 6(10):e2336483–e2336483, 2023.
- [5] Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hannah J. Hong, Mehr Kashyap, Adarsh R. Chaurasia, Nigam R. Shah, Karandeep Singh, Tina Tazbaz, Arnold Milstein, Marc A. Pfeffer, and Nigam H. Shah. Testing and evaluation of health care applications of large language models: a systematic review. *Jama*, 2025.
- [6] Hye Sun Yun and Timothy Bickmore. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. *Journal of Medical Internet Research*, 27:e68560, 2025.
- [7] Moustafa Laymouna, Yuanchao Ma, David Lessard, Tibor Schuster, Kim Engler, and Bertrand Lebouché. Roles, users, benefits, and limitations of chatbots in health care: rapid review. *Journal of medical Internet research*, 26:e56930, 2024.

- [8] Ryan T Hurt, Christopher R Stephenson, Elizabeth A Gilman, Christopher A Aakre, Ivana T Croghan, Manpreet S Mundi, Karthik Ghosh, and Jithinraj Edakkanambeth Varayil. The use of an artificial intelligence platform openevidence to augment clinical decision-making for primary care physicians. *Journal of Primary Care & Community Health*, 16:21501319251332215, 2025.
- [9] Jeanette C Prorok, Emma C Iserman, Nancy L Wilczynski, and Robert B Haynes. The quality, breadth, and timeliness of content updating vary substantially for 10 online medical texts: an analytic survey. *Journal of clinical epidemiology*, 65(12):1289–1295, 2012.
- [10] Isaac A. Bernstein, Youchen Victor Zhang, Devendra Govil, Iyad Majid, Robert T. Chang, Yang Sun, Ann Shue, Jonathan C. Chou, Emily Schehlein, Karen L. Christopher, Suzann L. Groth, Caitlin Ludwig, and Sophia Y. Wang. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA network open*, 6(8):e2330320–e2330320, 2023.
- [11] Bright Huo, Gary S. Collins, David Chartash, Arun James Thirunavukarasu, Annette Flanagan, Alfonso Iorio, Giovanni Cacciamani, Xi Chen, Nan Liu, Piyush Mathur, An-Wen Chan, Christine Laine, Daniela Pacella, Michael Berkwits, Stavros A. Antoniou, Jennifer C. Camaradou, Carolyn Canfield, Michael Mittelman, Timothy Feeney, Elizabeth Loder, Riaz Agha, Ashirbani Saha, Julio Mayol, Anthony Sunjaya, Hugh Harvey, Jeremy Y. Ng, Tyler McKechnie, Yung Lee, Nipun Verma, Gregor Stiglic, Melissa McCradden, Karim Ramji, Vanessa Boudreau, Monica Ortenzi, Joerg Meerpohl, Per Olav Vandvik, Thomas Agoritsas, Diana Samuel, Helen Frankish, Michael Anderson, Xiaomei Yao, Stacy Loeb, Cynthia Lokker, Xiaoxuan Liu, Eliseo Guallar, Gordon Guyatt, and The CHART Collaborative. Reporting guideline for chatbot health advice studies: the chart statement. *JAMA Network Open*, 8(8):e2530220–e2530220, 2025.
- [12] Ilana Goldowitz, Tequam Worku, Lisa Brown, and Harvey V. Fineberg, editors. *A Long COVID Definition: A Chronic, Systemic Disease State with Profound Consequences*. National Academies Press, Washington, DC, July 2024. National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Global Health; Board on Health Sciences Policy; Committee on Examining the Working Definition for Long COVID. PMID: 39110819.
- [13] Roger Chou, Eric Herman, Azrah Ahmed, Jordan Anderson, Shelley Selph, Tracy Dana, Leah Williams, and Ilya Ivlev. Long covid definitions and models of care: a scoping review. *Annals of internal medicine*, 177(7):929–940, 2024.
- [14] E Wesley Ely, Lisa M Brown, and Harvey V Fineberg. Long covid defined, 2024.
- [15] W Michael Brode and Esther Melamed. A practical framework for long covid treatment in primary care. *Life Sciences*, 354:122977, 2024.
- [16] Abby L. Cheng, Eric Herman, Benjamin Abramoff, Jordan R. Anderson, Alba Azola, John M. Baratta, Matthew N. Bartels, Ratna Bhavaraju-Sanka, Svetlana Blitshteyn, Jeffrey S. Fine, Talia K. Fleming, Monica Verduzco-Gutierrez, Janna E. Herrera, Ravi Karnik, Megan Kurylo, Matthew T. Longo, Molly D. McCauley, Esther Melamed, Mitchell G. Miglis, Janna D. Neal, C. Vaile Oleson, David Putrino, Lindsay Rydberg, Julie K. Silver, Carina M. Terzic, Joanne H. Whiteson, and William N. Niehaus. Multidisciplinary collaborative guidance on the assessment and treatment of patients with long covid: A compendium statement. *PM&R*, 17(6):684–708, 2025.
- [17] Ziyad Al-Aly, Hannah Davis, Lisa McCorkell, Letícia Soares, Sarah Wulf-Hanson, Akiko Iwasaki, and Eric J Topol. Long covid science, research and policy. *Nature medicine*, 30(8):2148–2164, 2024.
- [18] C Jessica E Metcalf and Justin Lessler. Opportunities and challenges in modeling emerging infectious diseases. *Science*, 357(6347):149–152, 2017.
- [19] John Zarocostas. How to fight an infodemic. *Lancet*, 395(10225):676, 2020.
- [20] Jie Song, Zhichuan Xu, Mengqiao He, Jinhua Feng, and Bairong Shen. Graph retrieval augmented large language models for facial phenotype associated rare genetic disease. *NPJ Digital Medicine*, 8(1):543, 2025.
- [21] Charlotte Zelin, Wendy K Chung, Mederic Jeanne, Gongbo Zhang, and Chunhua Weng. Rare disease diagnosis using knowledge guided retrieval augmentation for chatgpt. *Journal of Biomedical Informatics*, 157:104702, 2024.
- [22] Hai Li, Jingyi Huang, Mengmeng Ji, Yuyi Yang, and Ruopeng An. Use of retrieval-augmented large language model for covid-19 fact-checking: Development and usability study. *Journal of Medical Internet Research*, 27:e66098, 2025.

- [23] Akihiko Wada, Yuya Tanaka, Mitsuo Nishizawa, Akira Yamamoto, Toshiaki Akashi, Akifumi Hagiwara, Yayoi Hayakawa, Junko Kikuta, Keigo Shimoji, Katsuhiro Sano, Koji Kamagata, Atsushi Nakanishi, and Shigeki Aoki. Retrieval-augmented generation elevates local llm quality in radiology contrast media consultation. *NPJ Digital Medicine*, 8(1):395, 2025.
- [24] Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, Shu-Chun Wu, Vesela P. Kovacheva, and Daniel Shu Wei Ting. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *NPJ Digital Medicine*, 8(1):187, 2025.
- [25] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- [26] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.
- [27] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.
- [28] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *International Conference on Learning Representations*, 2021.
- [29] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.
- [30] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, 2024.
- [31] Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. A systematic review of large language model (llm) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25:1–19, 2025. doi: 10.1186/s12911-025-02954-4.
- [32] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi, Asad Aali, Ashwin Nayak, Shivam Vedak, Sneha S. Jain, Birju Patel, Oluseyi Fayanju, Shreya Shah, Ethan Goh, Dong-han Yao, Brian Soetikno, Eduardo Reis, Sergios Gatidis, Vasu Divi, Robson Capasso, Rachna Saralkar, Chia-Chun Chiang, Jenelle Jindal, Tho Pham, Faraz Ghoddusi, Steven Lin, Albert S. Chiou, Christy Hong, Mohana Roy, Michael F. Gensheimer, Hinesh Patel, Kevin Schulman, Dev Dash, Danton Char, Lance Downing, Francois Grolleau, Kameron Black, Bethel Mieso, Aydin Zahedivash, Wen-wai Yim, Harshita Sharma, Tony Lee, Hannah Kirsch, Jennifer Lee, Nerissa Ambers, Carlene Lugtu, Aditya Sharma, Bilal Mawji, Alex Alekseyev, Vicky Zhou, Vikas Kakkar, Jarrod Helzer, Anurang Revri, Yair Barnett, Roxana Daneshjou, Jonathan Chen, Emily Alsentzer, Keith Morse, Nirmal Ravi, Nima Aghaeepour, Vanessa Kennedy, Akshay Chaudhari, Thomas Wang, Sanmi Koyejo, Matthew P. Lungren, Eric Horvitz, Percy Liang, Mike Pfeffer, and Nigam H. Shah. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025.
- [33] Paul Hager, Friederike Jungmann, Robbie Holland, Kshitij Bhagat, Ivy Hubrecht, Marcus Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9): 2613–2622, 2024. doi: 10.1038/s41591-024-03097-1.
- [34] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623, 2023.
- [35] Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*, 2024.



- [36] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, 2024.
- [37] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitan, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.
- [38] Jaime A Teixeira da Silva, Panagiotis Tsigaris, and Mohammadamin Erfanmanesh. Publishing volumes in major databases related to covid-19. *Scientometrics*, 126:831–842, 2021. doi: 10.1007/s11192-020-03675-3.
- [39] Nicolas Fiorini, Kathi Canese, Grisha Starchenko, Evan Kireev, Won Kim, Vicky Miller, Mike Osipov, Mark Kholodov, Rafis Ismagilov, Srinivasan Mohan, James Ostell, and Zhiyong Lu. Best match: New relevance search for pubmed. *PLoS Biology*, 16(8):e2005343, August 2018. doi: 10.1371/journal.pbio.2005343. PMID: 30153250; PMCID: PMC6112631.

## Appendix A: Prompts

### Common generation prompt

The following prompt is used for generation of answers in all corpus selections.

You are a medical chatbot that is answering a long COVID question asked by a clinician.  
Format your response using markdown with section headers and bulleted lists.

### Inline citation prompt

The following prompt is used to generate a response to the user question based on the document chunks retrieved. This only applies to the G-1, GS-4, R-110, and PM modes.

Use the provided documents to answer the question. Create inline citations by providing the document's "short\_citation" string surrounded by parentheses to show exactly where each line of the response originates from. Only incorporate evidence from the provided documents that is relevant to answering the question.

### Web search prompt

The following prompt is used for GPT-4o when web search is enabled.

You must reply using information from peer reviewed medical publications only. Please cite your sources.

### PubMed Query Prompt

The following prompt takes the user question and creates a PubMed query using a single-shot example.

You are a PubMed search expert. Create a PubMed advanced search query for the following question:

Question: {question}

Requirements:

1. MUST include 'AND "long COVID"' in the query
2. Use PubMed advanced search syntax with parentheses, AND, OR operators
3. Return only the search query, nothing else

Examples:

Question: "What are the specific neurological symptoms of long COVID in elderly patients?"

Query: (neurological OR neurology) AND (elderly OR seniors) AND "long COVID"

## PubMed “Rewrite Query” Prompt

When insufficient results are returned, this prompt is used to create a new PubMed search query that seeks to elicit more results than the previous query.

You are a PubMed search expert. The previous search query failed to find enough results. Create a more relaxed PubMed search query based on the original question.

Original Question: {question}

Failed Query: {failed\_query}

Requirements for the relaxed query:

1. MUST include 'AND "long COVID"' in the query
2. Use more general terms and broader synonyms than the failed query
3. If the terms cannot be made more general, use fewer terms

## Pairwise Evaluation Prompt

The following prompt compares two responses head-to-head based on a given criteria. The “Context for Answer 1” and “Context for Answer 2” were included when evaluating faithfulness and overall. Context was not included when evaluating relevance and comprehensiveness. This prompt was adapted from Edge et al. [37] and Zheng et al. [34].

—Goal— Given a question and two answers (Answer 1 and Answer 2), assess which answer is better according to the following measure: {criteria\_text}

Your assessment should include two parts (in the given order):

Reasoning: a short explanation of why you chose the winner with respect to the measure described above.

Winner: this should be 1 (if Answer 1 is better), 2 (if Answer 2 is better), or 0 if they are fundamentally similar and the differences are immaterial.

Ensure that the order in which the responses were presented does not influence your decision.

Do not allow the length of the responses to influence your decision.

Question: {question}

Context for Answer 1: {retrieved chunks for answer 1}

Context for Answer 2: {retrieved chunks for answer 2}

Answer 1: {answer 1}

Answer 2: {answer 2}

Assessment:

## Output format prompt

The following was used with the pairwise evaluation prompt to ensure the output would be json that could be parsed.

IMPORTANT: Respond with ONLY a valid JSON object in exactly this format:

```
{“comparisons”: [{“reasoning”: “explanation for this specific comparison”, “winner”: Z, }, ...]}
```

Where:

- reasoning is a short explanation specific to this pairwise comparison
- Z is the winner (1, 2, or 0 for tie)

Do not include any text before or after the JSON. Do not use markdown formatting or code blocks.

## Faithfulness criteria

The following was adapted from Edge et al. [37] and used with the pairwise evaluation prompt as the criteria for faithfulness.

Is the answer supported by its associated context? If the answer contains claims that cannot be substantiated by the retrieved context, it is unfaithful, even if it is factually correct. For example, the question is “What are the benefits and drawbacks of nuclear energy”, and the retrieved context involves information from a study that shows radiation released during the

Three Mile Island accident is related to increased cancer incidence in the area. A faithful answer would cite the Three Mile Island accident to warn of its potential danger. An answer citing the Chernobyl accident, without being present in the retrieved context, would be unfaithful. An answer that incorrectly uses the Three Mile Island accident to say that nuclear energy does not pose potential risks would be unfaithful.

### **Relevance criteria**

The following was adapted from Edge et al. [37] and used with the pairwise evaluation prompt as the criteria for relevance.

Does the answer provide information that relates to the question? If the answer contains digressions from topic established by the question, it would demonstrate irrelevance. For example, if the question is “What are the benefits and drawbacks of nuclear energy”, a relevant answer may discuss the large number of kilowatts it produces annually, as well as the risks posed by a nuclear meltdown. An answer that discusses the danger of a nuclear bomb, without connecting this to nuclear power generation, would be irrelevant.

### **Comprehensiveness criteria**

The following was adapted from Edge et al. [37] and used with the pairwise evaluation prompt as the criteria for comprehensiveness.

How much detail does the answer provide to cover all the aspects of the question? A comprehensive answer should be thorough and complete. For example, if the question is “What are the benefits and drawbacks of nuclear energy”, a comprehensive answer would provide both the positive and negative aspects of nuclear energy, such as its efficiency, environmental impact, safety, cost, etc. A comprehensive answer should not leave out any important points. For example, an incomplete answer would only provide the benefits of nuclear energy without describing the drawbacks.

### **Overall criteria**

The following was used with the pairwise evaluation prompt as the criteria for the overall metric.

Each of the following criteria should be weighted equally. Faithfulness: the answer should be supported by its associated context. If the answer contains claims that cannot be substantiated by the retrieved context, it is unfaithful, even if it is factually correct. Comprehensiveness: the answer should provide details to cover all the aspects of the question. Relevance: the answer should provide information that relates to the question without digressions.

## **Appendix B: LongCOVID-CQ dataset**

### **General Knowledge and Evidence Synthesis**

1. What are the most common symptoms of Long COVID, and how do they vary between patients?
2. Can you summarize the current evidence on the pathophysiology of Long COVID?
3. What role does autonomic dysfunction play in Long COVID, and what are the management strategies?
4. What evidence-based therapies exist for treating fatigue in Long COVID?
5. What does the current evidence say about functional impairment and disability in patients with Long COVID, and how can patients get disability protections?

### **Clinical Scenarios and Decision Support**

1. A 45-year-old patient presents with persistent shortness of breath 8 months after mild COVID-19. What tests would you recommend and why?
2. How would you differentiate Long COVID-related cognitive dysfunction from other causes of cognitive impairment?
3. What treatment options would you recommend for a patient with POTS (postural orthostatic tachycardia syndrome) related to Long COVID?

4. A patient with Long COVID presents with significant mental health symptoms. What screening tools and management steps should be considered?
5. How would you approach rehabilitation for a patient with Long COVID-related post-exertional malaise (PEM)?
6. A 72 year-old black woman with pre-existing hypothyroidism and fibromyalgia presents with severe fatigue, muscle aches, chest pain, and abdominal pain following COVID-19 infection in July 2022. What tests should be ordered, and what treatment should she receive?

### Controversies and Evolving Evidence

12. What is the current evidence for antiviral therapy or immunomodulators in managing Long COVID?
13. How reliable is the use of biomarkers, such as CRP or cytokines, in diagnosing or monitoring Long COVID?
14. What does the evidence say about the impact of COVID-19 vaccination on Long COVID symptoms?
15. Is there any evidence to support dietary interventions or supplements for Long COVID management?
16. What does current evidence say about the duration of Long COVID symptoms, and are there predictors of recovery?

### Safety and Limitations

17. What are the potential pitfalls in diagnosing Long COVID in a patient with multiple comorbidities?
18. What are the limitations of current evidence on Long COVID treatment?
19. How do you ensure patients with Long COVID receive equitable care?
20. Patients are using nicotine patches to treat Long COVID, is that an effective treatment? Why would nicotine improve Long COVID symptoms?

## Appendix C: Retrieval architecture

**Dense retrieval implementation.** For expert-curated corpora (G-1, GS-4, R-110), we implemented a preprocessing pipeline for full-text document access and vector database construction. Full texts were retrieved via the PubMed Central (PMC) API when available; otherwise, PDFs were converted to plaintext using PyPDF with manual removal of non-content elements (headers, metadata, figures, numerical tables). Documents were segmented using recursive character splitting with 1200-character chunks and 600-character overlap to preserve paragraph structure. Each chunk was embedded using OpenAI’s text-embedding-3-small model (dimensionality: 1536) and indexed in a FAISS vector database. During inference, queries were embedded using the same model, and the top-25 nearest neighbors (cosine similarity) were retrieved and concatenated with citation metadata for response generation via LangChain and GPT-4o.

**Hybrid sparse-dense retrieval for PubMed.** The PM corpus employed a two-stage retrieval process. Initially, GPT-4o generated PubMed search queries from input questions using single-shot prompting. PubMed’s Best Match algorithm (BM25 + LambdaMART ranking) returned candidate documents [39]. If fewer than 25 results were obtained, query generalization was iteratively applied until the threshold was met (Appendix A). Full texts were retrieved via the PMC API (abstracts used when unavailable), then processed through the same chunking and dense retrieval pipeline described above, yielding the top-25 most relevant chunks for generation.

## Appendix D: GS-4 response example

We show an example response to the question “What role does autonomic dysfunction play in Long COVID, and what are the management strategies?” using the GS-4 corpus.

### Role of Autonomic Dysfunction in Long COVID

Autonomic dysfunction is a significant component of Long COVID, affecting a substantial number of patients. Studies have shown that a large proportion of individuals with Long COVID experience moderate to severe autonomic dysfunction, as indicated by high scores on the Composite Autonomic Symptom 31 (COMPASS-31) assessment (Fineberg\_39110819). Common manifestations include postural orthostatic tachycardia syndrome (POTS) and orthostatic intolerance, which can significantly impact daily functioning (Fineberg\_39110819).

**Question:** What role does autonomic dysfunction play in Long COVID, and what are the management strategies?

GS-4 response Guideline + Systematic reviews	R-110 response References of the guideline	PM response PubMed
<b>Role of Autonomic Dysfunction in Long COVID</b>  <b>Management Strategies</b> <b>Non-Pharmacologic Management</b> <ul style="list-style-type: none"> <li>•Hydration and Salt Intake</li> <li>•Compression Garments</li> <li>•Physical Counterpressure Maneuvers</li> <li>•Autonomic Rehabilitation</li> </ul> <b>Pharmacologic Management</b> <ul style="list-style-type: none"> <li>•Medication Review</li> <li>•Severe Symptoms</li> </ul> <b>Referral and Specialist Consultation</b> <ul style="list-style-type: none"> <li>•Autonomic Specialist</li> <li>•Cardiologist</li> </ul> <b>Holistic Management</b> <ul style="list-style-type: none"> <li>•Validation and Support</li> <li>•Tailored Physical Activity</li> </ul>	<b>Role of Autonomic Dysfunction in Long COVID</b>  <b>Pathophysiology</b> <ul style="list-style-type: none"> <li>•Sympathetic Overactivation</li> <li>•Underactive Vagus Nerve</li> </ul> <b>Management Strategies for Autonomic Dysfunction in Long COVID</b> <b>Non-Pharmacological Interventions</b> <ul style="list-style-type: none"> <li>•Rehabilitation Programs</li> <li>•Self-Management Techniques</li> </ul> <b>Pharmacological Interventions</b> <ul style="list-style-type: none"> <li>•Stellate Ganglion Block (SGB)</li> </ul> <b>Research and Future Directions</b> <b>Considerations</b> <ul style="list-style-type: none"> <li>•Health Equity</li> <li>•Individualized Care</li> </ul>	<b>Role of Autonomic Dysfunction in Long COVID</b>  <b>Pathophysiology</b> <ul style="list-style-type: none"> <li>•Sympathetic Overactivation</li> <li>•Inflammation</li> <li>•Neurological Contributions</li> </ul> <b>Management Strategies</b>  <b>Non-Pharmacological Interventions</b> <ul style="list-style-type: none"> <li>•Exercise Training</li> <li>•Vagus Nerve Stimulation</li> </ul> <b>Pharmacological Interventions</b> <ul style="list-style-type: none"> <li>•Autonomic Modulating Drugs</li> <li>•Sympathetic Response Mitigation</li> </ul> <b>Multidisciplinary Approach</b> <ul style="list-style-type: none"> <li>•Cardiovascular Rehabilitation Programs (CRPs)</li> <li>•Integrated Care Pathways</li> </ul> <b>Monitoring and Diagnosis</b> <ul style="list-style-type: none"> <li>•Heart Rate Variability (HRV)</li> </ul>

Figure 3: Comparison of answers to an example question from the LongCOVID-CQ dataset using the GS-4, R-110, and PM corpuses. Each response's section headers are included in the figure for conciseness.

## Management Strategies

### Non-Pharmacologic Management

- **Hydration and Salt Intake:** Increase oral intake to 2.5–3.5 liters of fluid and 2.8–4 grams of sodium daily (AAPMRCompendium\_NP002).
- **Compression Garments:** Use waist-high stockings and/or abdominal binders to help manage orthostatic intolerance (AAPMRCompendium\_NP002).
- **Physical Counterpressure Maneuvers:** Techniques such as standing with legs crossed or sitting in a knee-chest position can mitigate symptoms (AAPMRCompendium\_NP002).
- **Autonomic Rehabilitation:** Engage in a personalized rehabilitation program, starting with supine core strengthening or recumbent activities like biking or swimming (AAPMRCompendium\_NP002).

### Pharmacologic Management

- **Medication Review:** Consider weaning or de-prescribing medications that may exacerbate symptoms such as orthostatic intolerance, tachycardia, and hypotension (AAPMRCompendium\_NP002).
- **Severe Symptoms:** Medication management should be considered if symptoms are severe or persistent after non-pharmacologic measures. Intravenous fluids may be used sparingly during symptom flares (AAPMRCompendium\_NP002).

### Referral and Specialist Consultation

- **Autonomic Specialist:** Referral to an autonomic specialist is recommended if non-pharmacologic and first-line pharmacologic management do not ameliorate symptoms, or if there is frequent syncope or a need for additional autonomic function assessment (AAPMRCompendium\_NP002).
- **Cardiologist:** If an autonomic specialist is unavailable, a general cardiologist may be consulted for management outside the scope of the treating clinician (AAPMRCompendium\_NP002).

**Holistic Management** Validation and Support: It is crucial to validate the patient's experience and provide reassurance about the reality of Long COVID symptoms, emphasizing that many patients can manage their symptoms and improve their quality of life with available treatment strategies (AAPMRCompendium\_NP002).

Tailored Physical Activity: Physical activity recommendations should be carefully tailored to avoid triggering post-exertional malaise (AAPMRCompendium\_NP002).

These strategies aim to address the complex and multifaceted nature of autonomic dysfunction in Long COVID, providing a comprehensive approach to management.

## Appendix E: R-110 response example

We show an example response to the question “What role does autonomic dysfunction play in Long COVID, and what are the management strategies?” using the R-110 corpus.

### Role of Autonomic Dysfunction in Long COVID

Autonomic dysfunction, particularly dysautonomia, plays a significant role in the pathophysiology of Long COVID. The autonomic nervous system (ANS) is crucial in managing acute hyperinflammation and endothelial dysfunction during viral infections like COVID-19. In Long COVID, a malfunctioning ANS is often marked by sympathetic hyperactivity and an underactive vagus nerve, which may lead to persistent inflammation and disrupt the balance between the sympathetic and parasympathetic systems (Pearson\_37711269). This imbalance is thought to contribute to the characteristic symptoms of Long COVID, such as fatigue, cognitive issues, and cardiovascular dysautonomia (Rubin\_32965460).

### Management Strategies for Autonomic Dysfunction in Long COVID

#### Non-Pharmacological Interventions

- **Rehabilitation Programs:** Cardiovascular rehabilitation and telerehabilitation exercise programs have been shown to improve cognitive deficits in Long COVID patients (Sathyamoorthy\_38206585).
- **Self-Management Techniques:** Patients often resort to self-prescribed modalities, including dietary modifications and over-the-counter supplements. However, these should be pursued with caution and under professional guidance to avoid risks such as drug interactions (Sweis\_38541324).

**Pharmacological Interventions** **Stellate Ganglion Block (SGB):** This intervention targets the sympathetic nervous system and has shown promise in alleviating symptoms of dysautonomia in Long COVID by increasing cerebral and regional blood flow, potentially resetting ANS function (Khan\_36628048).

**Research and Future Directions** The PASC Collaborative highlights the need for further studies to determine biomarkers and appropriate diagnostic and therapeutic approaches for post-COVID autonomic dysfunction (Blitshteyn\_36169154).

Clinical trials are ongoing to evaluate the effectiveness of various pharmacological interventions, including immunotherapy and other novel treatments (Reis\_NP020).

#### Considerations

- **Health Equity:** Special considerations are necessary for populations with pre-existing autonomic disorders and those facing health disparities, as these groups may experience a higher burden of Long COVID symptoms (Blitshteyn\_36169154).
- **Individualized Care:** Given the variability in symptoms and responses to treatment, personalized management plans are essential. This includes tailored investigations and specialized consultations based on individual symptoms (Sweis\_38541324).

In summary, autonomic dysfunction is a key component of Long COVID, and while management strategies are still evolving, a combination of rehabilitation, self-management, and targeted pharmacological interventions shows potential. Ongoing research is crucial to develop evidence-based guidelines and improve patient outcomes.

## Appendix F: PM response example

We show an example response to the question “What role does autonomic dysfunction play in Long COVID, and what are the management strategies?” using the PM corpus.

### Role of Autonomic Dysfunction in Long COVID

Autonomic dysfunction is a significant component of Long COVID, manifesting as dysautonomia, which includes symptoms such as fatigue, arrhythmias, and hypertension. This dysfunction is linked to the cardiovascular and neurological aspects of Long COVID (Allendes\_36743679). Dysautonomia in Long COVID can present as postural orthostatic tachycardia syndrome (POTS), characterized by an exaggerated sympathetic response and orthostatic intolerance (Astin\_36412084; Ewing\_40254579).

## Pathophysiology

- **Sympathetic Overactivation:** Chronic activation of the sympathetic nervous system can lead to arrhythmogenesis, orthostatic hypotension, and cognitive decline (Allendes\_36743679).
- **Inflammation:** Microinflammation in autonomic nerve fibers, termed “post-COVID Guillan-Barré syndrome” (PCGBS), is a specific type of dysautonomia observed in Long COVID (Allendes\_36743679).
- **Neurological Contributions:** Dysautonomia and brain fog suggest neurological involvement, with inflammation playing a critical role in the pathogenesis (Astin\_36412084).

## Management Strategies

### Non-Pharmacological Interventions

- **Exercise Training:** Exercise has shown potential in improving cardiovascular and autonomic function. It may enhance vascular endothelial function and autonomic balance, making it a promising complementary strategy for Long COVID patients with dysautonomia (Allendes\_36743679).
- **Vagus Nerve Stimulation:** Techniques that stimulate the vagus nerve, such as breath retraining, HRV biofeedback, and yoga, can help reduce sympathetic activity and increase vagal tone (Astin\_36412084; Ewing\_40254579).

### Pharmacological Interventions

- **Autonomic Modulating Drugs:** Medications like ivabradine, beta-blockers, and midodrine can be useful in managing symptoms of dysautonomia in Long COVID patients (Ewing\_40254579).
- **Sympathetic Response Mitigation:** beta-receptor blockers may help mitigate symptoms from exaggerated sympathetic responses (Astin\_36412084).

### Multidisciplinary Approach

- **Cardiovascular Rehabilitation Programs (CRPs):** These programs are interdisciplinary interventions that can stimulate the parasympathetic nervous system, improve cardiorespiratory fitness, and reduce cardiovascular risk factors (Allendes\_36743679).
- **Integrated Care Pathways:** Combining rehabilitation, drug interventions, and research into healthcare access can provide comprehensive management for Long COVID (Astin\_36412084).

**Monitoring and Diagnosis Heart Rate Variability (HRV):** HRV analysis is a suitable tool for diagnosing post-COVID dysautonomia and guiding treatment strategies (Allendes\_36743679).

In summary, autonomic dysfunction plays a crucial role in Long COVID, with management strategies focusing on both non-pharmacological and pharmacological interventions to restore autonomic balance and improve patient outcomes.

## Appendix G: LLM-as-a-judge comparison

The following was the output of the LLM-as-a-judge based on comprehensiveness between GS-4 (1) vs. R-110 (2).

Answer 1 provides a more comprehensive overview of both the role of autonomic dysfunction in Long COVID and the management strategies. It details specific non-pharmacologic and pharmacologic interventions, including hydration, compression garments, and medication review, as well as the importance of specialist consultation and holistic management. Answer 2, while informative, focuses more on the theoretical aspects and future research directions, lacking the detailed management strategies provided in Answer 1.

## Appendix H: R-110 Corpus Preprocessing

Two references of the AAPM&R compendium were excluded from the R-110 corpus due to infeasibility of preprocessing. Reference 48's full text was not available in English, only in German. Reference 71 was a website compiling links to social security regulations, which would require comprehensive web-scraping to parse with limited relevance to our clinical questions. Both citations are shown below.

48 Laskowski NM, Brandt G, Paslakis G. Geschlechtsspezifische Unterschiede und Ungleichheiten der COVID-19 Pandemie: Eine Synthese systematischer Reviews unter Einbeziehung sexueller und geschlechtlicher Minderheiten. *Psychother Psychosom Med Psychol.* 2024;74(2):57-69. doi:10.1055/a-2228-6244

71 Social Security Administration. Code of Federal Regulations: PART 404 FEDERAL OLD-AGE, SURVIVORS AND DISABILITY INSURANCE (1950); U.S. Social Security Administration, Revised as of April 1, 2023; accessed via [https://www.ssa.gov/OP\\_Home/cfr20/404/404-0000.htm](https://www.ssa.gov/OP_Home/cfr20/404/404-0000.htm) on August 4, 2024.

At the time our research was conducted, the latest version of the AAPM&R compendium had 113 references total. Future published versions may have a different number of references due to subsequent author revisions.