PATHFinder Agent for Tailored Prenatal Care

Vaibhav Balloli
University of Michigan
Ann Arbor, USA
vballoli@umich.edu

Carissa Samuel
University of Michigan
Ann Arbor, USA

Samia Abdelnabi
University of Michigan
Ann Arbor, USA

Alex Peahl
University of Michigan
Ann Arbor, USA

Elizabeth Bondi-Kelly
University of Michigan
Ann Arbor, USA
ecbk@umich.edu

Abstract

Prenatal care is an important preventive service designed to improve outcomes for pregnant individuals. The American College of Obstetricians and Gynecologists (ACOG) recently introduced guidelines advocating tailored prenatal care, called PATH (Plan for Tailored Healthcare). We present `PATHFinder Agent` (Planner for Appropriate Tailored Healthcare), an end-to-end conversational agentic system that gathers patient health and social context through structured dialogue, curates individualized prenatal care plans aligned with PATH guidelines, and surfaces community resources from Michigan 211. The system features a four-stage workflow spanning patient intake, dynamic interaction, plan synthesis, and clinician oversight. We evaluate frontier large language models (LLMs) on expert-curated rubrics across five clinical dimensions, finding that GPT-5.2 achieves the highest average score (77.6%) while identifying key gaps in antenatal testing recommendations. We discuss future validation through human participant studies and randomized controlled trials.

CCS Concepts

• **Computing methodologies** → **Natural language processing**;
• **Human-centered computing** → **Natural language interfaces**;
• **Applied computing** → **Consumer health**.

Keywords

Large Language Models, Large Language Model Agents, Health, Reproductive Health, Maternal Health, Healthcare Agents, LLM-in-the-loop

ACM Reference Format:
Vaibhav Balloli, Carissa Samuel, Samia Abdelnabi, Alex Peahl, and Elizabeth Bondi-Kelly. 2026. PATHFinder Agent for Tailored Prenatal Care. In *Interactive Health Conference (IH '26), July 05–08, 2026, Porto, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3786579.3804996

1 Introduction

Prenatal care is a crucial preventive service that improves pregnancy outcomes for mothers and their children [23], with nearly four million pregnant patients each year in the United States receiving prenatal care. Prenatal care is multifaceted, involving planning and providing medical care, screening tests, answering questions, and connecting people to appropriate social and community resources.

The American College of Obstetricians and Gynecologists (ACOG), a professional association of physicians specializing in obstetrics and gynecology in the United States, recently proposed a transformation to existing guidelines, with the aim to begin **"carefully tailoring prenatal care"** [19]. These guidelines, called PATH (Plan for Appropriate Tailored Healthcare), broadly tackle **(a) addressing unmet social needs** and **(b) incorporating alternative care modalities** to help tailor the plan to the patient and provide prenatal care to many more birthing people. PATH was carefully designed after conducting interviews with 110 patients, clinicians, and policy makers representing 25 organizations and more than 75 clinics.

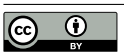
Adopting PATH is a complex task for both the patients and clinicians (Figure 1a). For patients with unmet social needs, additional prenatal visits with a maternity care professional are unlikely to address underlying needs and may create additional burden [24]. Furthermore, increased demands on physicians and other health care professionals to address patients' unmet social needs with insufficient resources have been associated with burnout [28].

We propose to help mitigate these challenges by creating a conversational AI agent-based interface, `PATHFinder Agent`, with abilities to **(a) follow up** with questions and clarifications requiring medical and conversational knowledge, **(b) perform actions via tool-calls** (searching for resources, generating a report, etc.), and **(c) follow instructions** to provide a comprehensive plan. We further carefully work towards avoiding unintended behaviors, ensuring oversight and grounding, and testing for deployment, as state-of-the-art medical systems like g-AMIE have called for [31]. We anticipate `PATHFinder Agent` has the potential to support broad deployment of the PATH guidelines, thereby improving health and well-being for pregnant and birthing individuals.

2 Background

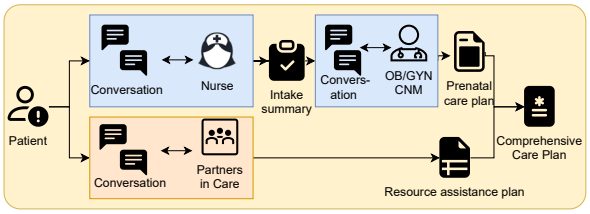
We identify three key areas from the literature that have particularly helped us shape the system.

Prenatal Care. Over 80% of adverse pregnancy outcomes are preventable through essential prenatal care and the management of unmet social needs [29]. However, the traditional 12–14 visit in-person model [11] is often inaccessible due to social drivers of

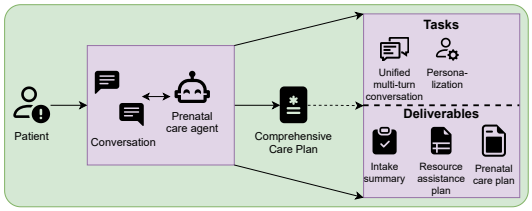


IH '26, Porto, Portugal

ACM ISBN 979-8-4007-2422-0/2026/07
https://doi.org/10.1145/3786579.3804996



(a) Current practices involve complex, disjoint conversations with patients to provide prenatal care.



(b) Centralized, scalable conversation with `PATHFinder Agent` equipped with the latest guidelines and grounded resources.

Figure 1: Illustrations of current practices (left) and our proposed solution (right)

health—such as housing instability and inflexible work schedules [14, 27]—leading many patients to be underserved and even feel unheard [4, 5, 13]. In response, ACOG developed the Plan for Appropriate Tailored Healthcare in pregnancy (PATH) [24], which utilizes shared decision-making to tailor care through flexible visit frequencies [1, 20, 30] and telehealth modalities [2], for example. Despite its potential, gaps remain in implementing PATH effectively within complex healthcare systems [15], and taking patient preferences into account will be vital [21, 22].

Medical Agents. Gemini, AMIE [10, 16, 26, 31] and MedAgentBench [9] focus on electronic health records, and MedAgentGym focuses on programming-centric agentic tasks in medicine[32]. These systems do not focus on integrating both domain-specific needs, like prenatal care, and domain-agnostic needs, like safety measures.

Tool-use and Conversational Systems. To utilize such medical agents to support patients and providers implementing PATH, we build on work where agents must ask clarifying questions and follow instructions to achieve the user's goal, as well as work in which agents need to call external tools. Research works like [18, 25] investigated LLMs' capabilities to invoke specific "tools" to solve tasks in mathematical reasoning, program synthesis, and general tasks. LLMs can decide to invoke a tool by generating tokens in an expected format, which enables them to navigate autonomously by recursively invoking tools. Instruction-tuned LLMs paired with strategies like ReACT (Reasoning and Acting) [34] have demonstrated near-perfect capabilities to accurately invoke tools. On the other hand, researchers have also studied task-oriented conversations [6, 7] between humans and automated evaluation of conversational models [8]. Subsequent works [3, 12, 17, 33] take a step closer to real-world applications with multi-turn interactions between the human and agents.

3 `PATHFinder Agent`: System Design

3.1 Problem Statement

Given a patient's medical characteristics and social context, `PATHFinder Agent` must (i) gather relevant information through structured, open-ended dialogue, (ii) curate an individualized prenatal care plan aligned with the specified guidelines, and (iii) surface Michigan 211 community resources matched to identified social needs—while enabling ongoing clinician review and oversight (Figure **??** illustrates the workflow).

Category	Tools
Medical	TOLAC Calculator; Defer to Clinician
Social needs resources	Get Groups; Get Categories; Get Subcategories; Get Resources by Subcategory & ZIP; No Resources Message
Personalization	Consult LLM for follow-up questions
Report	Add Patient Summary; Add Clinical Summary; Add Recommendations; Add Resources; Create Visit Schedule

Table 1: Agent tool suite (13 tools across 4 categories).

3.2 Objectives and Requirements

Core design requirements include interfaces and question flows that mirror existing clinical intake processes, and targeted, non-redundant follow-up questions with an interaction mode to prevent patient fatigue in long sessions while capturing data efficiently.

3.3 Design

Team. `PATHFinder` was co-designed by a board-certified Obstetrician and Gynecologist and Certified Nurse Midwife (CNM) alongside two computer scientists, ensuring jointly validated clinical and technical requirements.

Tools and Orchestration. The agent is equipped with 13 tools across four categories (Table1). Medical tools include a TOLAC (trial of labor after cesarean) calculator and a structured referral-to-clinician action. Resource tools implement a hierarchical Michigan 211 query interface. A personalization tool consults a secondary LLM to generate context-aware follow-up questions. Report tools incrementally build the patient and clinician summaries. The system instructions consists of ~14,000 tokens, with domain knowledge (from ACOG guidelines[19]), tool use instructions, workflow and safety policies, where the agent retrieves the knowledge and instructions to orchestrate the conversation.

Social needs resources data. We use Michigan 211[1] data organized by category (food, housing, transportation, utilities, clothing) and can be queried by subcategory and ZIP code via dedicated tool calls.

[1] https://mi211.org/

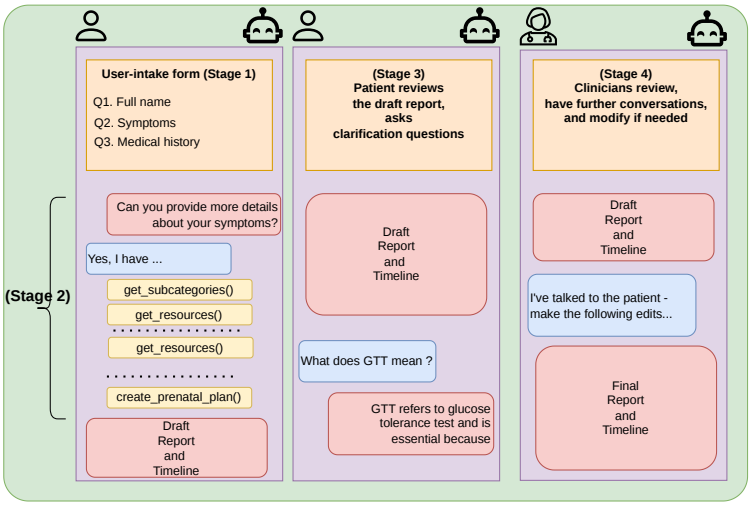


Figure 2: PATHFinder interfaces. `PATHFinder Agent` has a four-stage overflow: 1) Patient information intake (standardized form), 2) LLM-generated UI for structured, open-ended dialogue, 3) Draft report review and clarification for the patient and 4) Clinician review during provision of care.

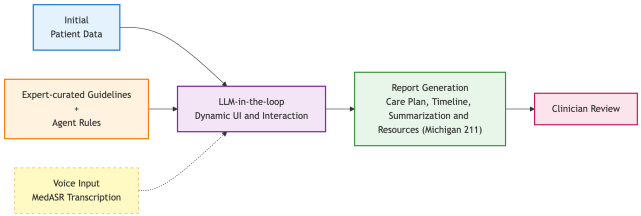


Figure 3: System architecture. React frontend communicates with a FastAPI backend routing to the `PATHFinder Agent` agent (LLM + tool executor), oversight classifier, FHIR service, and report generator. Conversation state is persisted in a relational database.

Interface design. `PATHFinder Agent` supports interaction via **Form+Chat Interface**, where the patient first completes a standardized intake form mirroring current clinical processes; the agent then transitions to an adaptive conversational phase to ask individualized follow-up questions, where another LLM reviews the questions and generates a form-based interface for the user to interact which supports buttons, check boxes, and switches to reduce the text entered by the user (see Figure 4b). Based on all the available user information and guidelines [19], appropriate reports and timelines are generated for the clinicians to review and patients to look at and clarify the content. Similarly, the agent is available for the clinician to make edits to the report if needed.

Workflow. The end-to-end workflow has four stages. *Stage 1 (Intake):* the patient fills a standardized form capturing demographics, medical history, gestational age, and social factors (Figure 4a). *Stage 2 (Dynamic interaction):* conditioned on responses and guidelines, the agent asks personalized follow-up questions to elicit unmet social needs, preferences, and barriers and uses tools in Table 1 to compose a plan (Figure 4b). *Stage 3 (Plan synthesis):* the agent invokes report tools to produce a patient-facing summary, a clinical summary, a visit-schedule recommendation, and curated Michigan 211 resources (Figure 4c). *Stage 4 (Clinician review):* the clinician reviews the plan with capabilities to approve or edit (Figure 4d).

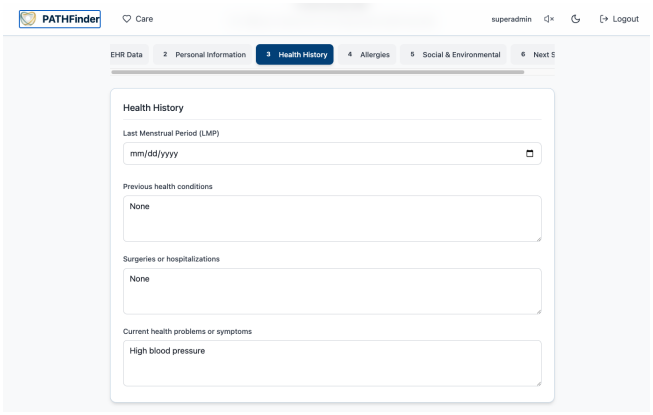
3.4 Clinician Oversight

Once the draft plan is generated, the clinician reviews the patient summary, report, and timeline to validate it and change it based on additional conversations and concerns. We highlight that this workflow potentially allows the clinician to better address concerns and provide care to patients while being involved in the prenatal care planning and validation of the plan.

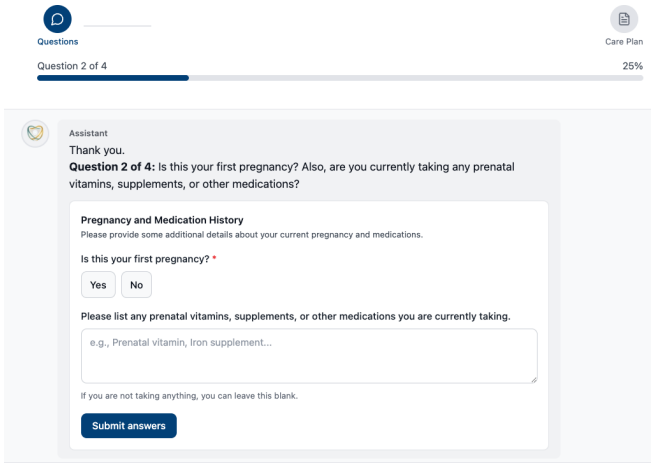
3.5 Evaluation

We evaluate `PATHFinder Agent` on synthetic patient profiles spanning diverse medical histories (e.g., high-risk pregnancies, TOLAC candidates). Each profile pairs with expert-curated rubrics specifying the expectations along the following dimensions: 1) the right visit frequency, 2) the right services (testing, recommendations, etc.), 3) timing for antenatal testing, 4) timing for growth ultrasound, and 5) modality of care to the patient (mandatory in-person, mix of in-person and group health, etc.).

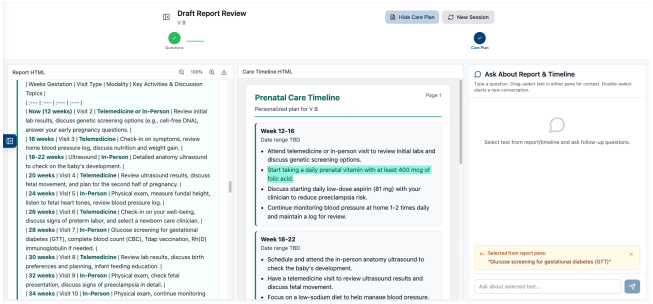
We adopt LLM-as-judge scoring to measure the performance of frontier models' recommendations across the five dimensions with the final score per condition normalized to 1. Table 5 shows aggregate rubric scores across state-of-the-art LLMs from OpenAI (GPT-5.2, GPT-4o) and Google (Gemini 2.5 pro and flash). Furthermore, Figure 5 breaks down scores by dimension, with visit frequency being the easiest across all models and recommending antenatal



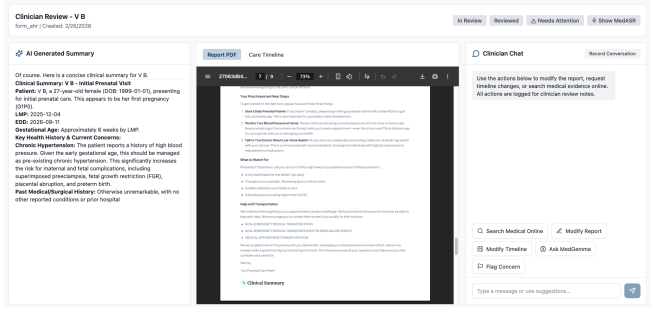
(a) Patient intake form (Stage1).



(b) Dynamic interaction with LLM generating intermediate UI.



(c) Patient draft report review and clarification (Stage 3).



(d) Clinician oversight and review (Stage4).

Figure 4: `PATHFinder Agent` interfaces. (a)Structured patient intake form; (b) `PATHFinder Agent` interviews the user to understand additional medical history details, and social determinants of health factors; (c) Patient reviews the draft report for transparency and clarification; (d)Clinician oversight dashboard showing conversation stage, risk flags, and escalation controls.

Model	Avg. Rubric Score(%)
GPT-5.2	77.60
Gemini 2.5 pro	71.50
Gemini 2.5 flash	62.00
GPT-4o	57.25

Table 2: Average rubric-based scores (0-100%) across frontier models. Higher is better.

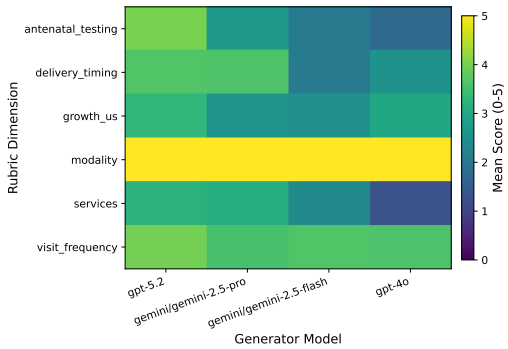


Figure 5: Per-dimension rubric scores. Antenatal testing and services recommendation show the widest performance gap across models (1 is lowest, 5 is highest).

testing and other services being the most difficult. These results suggest that robust oversight measures, both LLM- and human-driven, must be integrated into these systems with appropriate communication to ensure that deployed models do not make mistakes.

4 Future Work

Future research will focus on establishing formal accuracy guarantees for `PATHFinder Agent` to strengthen the system's technical reliability and performance standards. We will concurrently conduct a series of human participant experiments.

Acknowledgements

This work was partially supported by funding from Google and the University of Michigan (including the Raoul Wallenberg Institute,

E-Health and Artificial Intelligence, and the Center for Academic Innovation)

References

[1] E. Balk, V. Danilack, M. Bhuma, et al. 2023. Reduced Compared With Traditional Schedules for Routine Antenatal Visits: A Systematic Review. *Obstetrics & Gynecology* 142, 1 (2023), 8–18.

[2] E. Balk, V. Danilack, W. Cao, et al. 2023. Televisits Compared With In-Person Visits for Routine Antenatal Care: A Systematic Review. *Obstetrics & Gynecology* 142, 1 (2023), 19–29.

[3] Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. arXiv:2506.07982 [cs.AI] https://arxiv.org/abs/2506.07982

[4] M. Bellerose, M. Rodriguez, and P. Vivier. 2022. A systematic review of the qualitative literature on barriers to high-quality prenatal and postpartum care among low-income women. *Health Services Research* 57, 4 (2022), 775–785.

[5] M. Betron, T. McClair, S. Currie, and J. Banerjee. 2018. Expanding the agenda for addressing mistreatment in maternity care: a mapping review and gender analysis. *Reproductive Health* 15, 1 (2018), 143.

[6] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278* (2018).

[7] Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3002–3017.

[8] Izzeddin Gür, Dilek Hakkani-Tür, Gokhan Tür, and Pararth Shah. 2018. User Modeling for Task Oriented Dialogues. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. 900–906. doi:10.1109/SLT.2018.8639652

[9] Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. 2025. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. *NEJM AI* 2, 9 (2025), AIdbp2500144.

[10] Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

[11] S. Kilpatrick, L. Papile, and G. Macones. 2017. *Guidelines for Perinatal Care* (8th ed.). American Academy of Pediatrics/The American College of Obstetricians and Gynecologists, Elk Grove Village, IL/Washington, D.C.

[12] Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, et al. 2025. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 1160–1183.

[13] Y. Mohamoud, E. Cassidy, E. Fuchs, et al. 2023. Vital Signs: Maternity Care Experiences - United States, April 2023. *MMWR Morbidity and Mortality Weekly Report* 72, 35 (2023), 961–967.

[14] National Academies of Sciences, Engineering, and Medicine. 2019. *Integrating Social Care into the Delivery of Health Care: Moving Upstream to Improve the Nation's Health*. National Academies Press, Washington, D.C.

[15] M. Nijagal, D. Patel, C. Lyles, et al. 2021. Using human centered design to identify opportunities for reducing inequities in perinatal care. *BMC Health Services Research* 21, 1 (2021), 714.

[16] Anil Palepu, Valentin Liévin, Wei-Hung Weng, Khaled Saab, David Stutz, Yong Cheng, Kavita Kulkarni, S Sara Mahdavi, Joëlle Barral, Dale R Webster, et al. 2025. Towards conversational ai for disease management. *arXiv preprint arXiv:2503.06074* (2025).

[17] Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. The Berkeley Function Calling Leaderboard (BFCL): From Tool Use to Agentic Evaluation of Large Language Models. In *Forty-second International Conference on Machine Learning*.

[18] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2023. Gorilla: Large language model connected with massive apis, 2023. *URL https://arxiv.org/abs/2305.15334* (2023).

[19] Alex Peahl, Julia C Phillippi, and Mark A Turrentine. 2025. Tailored prenatal care delivery for pregnant individuals. *OBSTETRICS AND GYNECOLOGY* 145, 5 (2025), 565–577.

[20] A. F. Peahl, R. Gourevitch, E. Luo, et al. 2020. Right-Sizing Prenatal Care to Meet Patients' Needs and Improve Maternity Care Value. *Obstetrics & Gynecology* 135, 5 (2020), 1027–1037.

[21] A. F. Peahl, A. Novara, M. Heisler, V. K. Dalton, M. H. Moniz, and R. D. Smith. 2020. Patient Preferences for Prenatal and Postpartum Care Delivery: A Survey of Postpartum Women. *Obstetrics & Gynecology* 135, 5 (2020), 1038–1046.

[22] A. F. Peahl, A. Powell, H. Berlin, et al. 2021. Patient and provider perspectives of a new prenatal care model introduced in response to the coronavirus disease 2019 pandemic. *American Journal of Obstetrics and Gynecology* 224, 4 (2021), 384.e1–384.e11.

[23] Alex F Peahl, Roger D Smith, and Michelle H Moniz. 2020. Prenatal care redesign: creating flexible maternity care models through virtual care. *American journal of obstetrics and gynecology* 223, 3 (2020), 389–e1.

[24] A. F. Peahl, C. Zahn, M. Turrentine, et al. 2021. The Michigan Plan for Appropriate Tailored Health Care in Pregnancy Prenatal Care Recommendations. *Obstetrics & Gynecology* 138, 4 (2021), 593–602.

[25] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789* (2023).

[26] Khaled Saab, Jan Freyberg, Chunjong Park, Tim Strother, Yong Cheng, Wei-Hung Weng, David GT Barrett, David Stutz, Nenad Tomasev, Anil Palepu, et al. 2025. Advancing Conversational Diagnostic AI with Multimodal Reasoning. *arXiv preprint arXiv:2505.04653* (2025).

[27] J. Semega, M. Kollar, J. Creamer, and A. Mohanty. 2021. *Income and Poverty in the United States: 2018*. Technical Report. U.S. Census Bureau.

[28] Masami Tabata-Kelly, Xiaochu Hu, Michael J Dill, Philip M Alberti, Karen Bullock, William Crown, Malika Fair, Peter May, Pilar Ortega, and Jennifer Perloff. 2024. Physician engagement in addressing health-related social needs and burnout. *JAMA network open* 7, 12 (2024), e2452152–e2452152.

[29] S. Trost, J. Beauregard, G. Chandra, et al. 2022. Pregnancy-Related Deaths: Data from Maternal Mortality Review Committees in 36 US States, 2017–2019. https://www.cdc.gov/reproductivehealth/maternalmortality/erase-mm/data-mmrc.html.

[30] M. Turrentine. 2023. Prenatal Care Visit Frequency: How Much Is Too Much, and How Little Is Too Little? *Obstetrics & Gynecology* 142, 1 (2023), 6–7.

[31] Elahe Vedadi, David Barrett, Natalie Harris, Ellery Wulczyn, Shashir Reddy, Roma Ruparel, Mike Schaekermann, Tim Strother, Ryutaro Tanno, Yash Sharma, et al. 2025. Towards physician-centered oversight of conversational diagnostic AI. *arXiv preprint arXiv:2507.15743* (2025).

[32] Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Xiangru Tang, Hang Wu, May Dongmei Wang, Peifeng Ruan, Donghan Yang, Tao Wang, et al. 2025. Medagentgym: Training llm agents for code-based medical reasoning at scale. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

[33] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. arXiv:2406.12045 [cs.AI] https://arxiv.org/abs/2406.12045

[34] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.