Physician Perceptions of Large Language Models in Clinical Practice: A Mixed-Methods Survey Study

Francis Arellano[1*], Rohan Rani[1*], Jacqueline Sandling[1], Yumin Gao[1], Athena Nguyen[1], Jeya Anandakumar[1], Emily Chen[1], Deeya Garg[1], Enrico Bautista[1], Zahra Ahmad[1], Nika Shroff[1], Katherine Barnes[1,2], Joshua Biro[1,2], Kristen Miller[1,2]

**Co-first authors.*
[1] Georgetown University School of Medicine
[2] MedStar Health National Center for Human Factors in Healthcare
fca13@georgetown.edu, rmr137@georgetown.edu

Abstract

Large language models (LLMs) are rapidly emerging artificial intelligence (AI) tools with potential to transform clinical workflows, yet limited data exist on how physicians engage with them across specialties. We conducted a cross-sectional, mixed-methods survey of physicians across a large hospital system in the Northeastern United States between February and August 2025, using convenience and snowball sampling. The anonymous REDCap survey collected data surrounding demographics, AI familiarity, applications, documentation and communication practices, perceived challenges, and desired features. Quantitative data was analyzed descriptively and qualitative free-text responses thematically analyzed using Google Gemini 2.5 Pro. Fifty-two physicians participated, most commonly from internal medicine (32.7%) and surgery (26.9%), followed by pediatrics (17.3%) and emergency medicine (7.7%). Overall, 71.2% reported using at least one AI tool, most often ChatGPT. Among AI users (N=37), frequent applications included differential diagnosis (54.1%), research and data analysis (43.2%), documentation or administrative tasks (40.5%), and medical education (37.8%). Adoption patterns varied but did not significantly differ. Thematic analysis (N=36) identified four domains: AI as a documentation and communication assistant, knowledge synthesis and decision support, workflow optimization, and barriers & mistrust. While efficiency gains were reported, concerns regarding accuracy, medicolegal liability, bias, and lack of transparency were widespread, with nearly one-third citing insufficient training as a barrier. Physicians currently adopt LLMs pragmatically within a human-in-the-loop model, but safe and equitable integration will require regulatory clarity, AI literacy, and trust-building aligned with clinical needs.

1 Introduction

According to the American Medical Association's Augmented Intelligence Research, 68% of responding physicians cited AI as an advantage in patient care [1]. The integration of artificial intelligence (AI), specifically large language models (LLMs), into clinical practice is an area of rapid development and significant interest [2]. For example, AI-powered systems have been increasingly explored as a method to streamline and reduce the process of documentation [3]. Studies indicate that physicians may spend as much as two hours on electronic data entry for every hour of direct patient contact; this documentation burden is associated with increased medical errors and decreased job satisfaction [4]. The successful integration of LLMs into healthcare is contingent upon acceptance by frontline

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

clinicians [5]. Despite this, there are many barriers which continue to hinder widespread adoption of AI in clinical practice. Barriers include a lack of AI knowledge among the workforce, regulatory uncertainty, and concerns about workflow integration [5]. While a growing body of literature evaluates specialty-specific perceptions of LLMs, much of the existing literature on this topic is limited to single specialties [6, 7]. There is a relative scarcity of research that systematically investigates the perceptions, usage patterns, and concerns of a broad cross-section of practicing clinicians. To ensure that AI development aligns with clinical needs, it is necessary to understand the perspectives of its intended end-users. This study aims to address the existing gap by providing insight into physicians' current integration of LLMs in clinical practice, as well as compare perceived applications and barriers to integration across specialties including internal medicine and surgery.

2 Methods

This study employed a cross-sectional, mixed-methods survey design to investigate clinician perceptions and use of AI. A mixed-methods approach was chosen to integrate quantitative data with qualitative data. Data were collected between February 2025 and August 2025 via an anonymous, online survey created and disseminated using REDCap. Participants' survey data was deidentified and securely stored in a REDCap database, after which it was analyzed by trained research assistants. The target population comprised practicing physicians of a large hospital system in the Northeastern United States. The two primary recruitment sites were tertiary care academic hospitals. Participants were recruited through a non-probability, convenience sampling strategy using professional email listservs and social media networks, supplemented by snowball sampling. Of those who opened the survey website link, 47.8% completed a full response. No compensation was provided for completion of the survey.

The survey instrument was developed by the research team based on a review of existing literature and expert consultation. The survey was pilot-tested with five clinicians to assess clarity and completion time. The instrument was divided into three sections.

Section I: Demographics. This section collected quantitative data on participants' professional background, including degree, specialty, years in practice, and time allocation between clinical/research and inpatient/outpatient settings.

Section II: Experience with Artificial Intelligence Tools. This section used multiple-choice, "select all that apply," and ranked-choice questions to assess familiarity with specific LLMs, frequency of use, clinical applications, and perceived challenges. It also assessed practices around documenting and communicating AI use and identified features clinicians deemed most important in a medical AI tool.

Section III: Specialty Specific Questions. This section consisted of a single, mandatory, open-ended qualitative question: "What are some unique aspects of AI that you are using in your current specialty?"

Quantitative data were analyzed using SPSS Statistics for Windows, Version 28.0 (IBM Corp., Armonk, NY). Descriptive statistics (frequencies, percentages) were calculated for all categorical variables. To minimize missing data, key questions were mandatory and incomplete surveys were excluded from the final analysis. Qualitative data from Section III were analyzed by providing the labeled data to a publicly available LLM (Google Gemini, model 2.5 Pro Deep Research) to analyze for common themes.

3 Results

A total of 52 physicians (MD/DO) completed the survey and were included in the final analysis. The most represented specialties were internal medicine (N=17, 32.7%) and surgery (N=14, 26.9%). Our surgery cohort included a combination of general surgery (N=12), orthopedic surgery, and otolaryngology. There were 9 pediatricians (17.3%), 4 emergency medicine physicians (7.7%), and 8 other physicians split across adolescent medicine, anesthesiology, critical care, dermatology, neurology, obstetrics and gynecology, combined IM/pediatrics, and interventional radiology. Of the 52 respondents, 37 (71.2%) identified a primary AI tool they used in their practice. Of the 37 participants, 21 (57.0%) were still in residency or fellowship. Among the 16 attending physicians,

6 (16.2%) had been in practice for 0–5 years, 2 (5.4%) for 6–10 years, 6 (16.2%) for 11–20 years, and 2 (5.4%) for more than 20 years. Of the 15 respondents who did not report using an AI tool, 7 (46.7%) were still in residency or fellowship. Among the 8 attending physicians, 2 (13.3%) had been in practice for 0–5 years, 3 (20.0%) for 6–10 years, 2 (13.3%) for 11–20 years, and 1 (6.7%) for more than 20 years.

In terms of inpatient and outpatient practice settings (N=52), the most common arrangement was 25% outpatient/75% inpatient (N=17, 32.7%). Exclusive outpatient practice was reported by 17.3% (N=9), while exclusive inpatient practice was reported by 21.2% (N=11). Additionally, 21.2% (N=11) reported a 75% outpatient/25% inpatient split, and 7.7% (N=4) practiced with an even 50% outpatient/50% inpatient distribution. When asked about their clinical and research responsibilities (N=48), the majority reported practicing 100% clinical medicine (N=31, 64.5%). A substantial portion reported a 75% clinical/25% research split (N=12, 25.0%), while smaller groups reported a 50% clinical/50% research balance (N=2, 4.2%), a 25% clinical / 75% research distribution (N=2, 4.2%), or 100% research (N=1, 2.1%).

3.1 Quantitative Findings: AI Familiarity and Perceptions

Among the 52 survey respondents, ChatGPT was the most widely recognized tool, with 71.2% of all respondents reporting being "Familiar" or "Very Familiar." Familiarity with other tools was lower: Perplexity 15.4%, Claude 5.8%, and Microsoft Copilot 11.5%. A total of 37 respondents identified a primary AI tool they used in their practice. Reported frequency of use varied among these users, with 27.0% using their specified tool "daily" and 29.7% using the tool "occasionally."

When all respondents (N=52) were asked about the most important features in a future medical AI tool, "medical education and training" (48.1%) and "administrative tasks" (48.1%) were the most selected features, followed by "patient education" (46.3%).

Among responding clinicians who use an AI tool, the most common application was for "differential diagnosis generation" (54.1%), followed by "research and data analysis" (43.2%) and "medical education and training" (37.8%). More than half of clinicians who use AI (55.3%) reported that they do not formally document their use of the tools. When communication about AI use does occur, it is most often with "physicians and other colleagues" (60.5%) rather than "patients" (15.8%).

When stratifying our data into IM vs. Surgery vs. Other, we found no statistically significant difference in subjective use cases across specialties (Table 4). IM physicians reported the highest frequency of using AI for "differential diagnosis generation" (64.3%) and "medical education and training" (50.0%). In contrast, surgical respondents most often used AI for "research and data analysis" (60.0%), while clinicians in the "Other" category, including pediatrics, emergency medicine, and neurology, reported the greatest use for "documentation" (46.2%) and "administrative tasks" (30.8%). "Patient education" was most frequently cited by IM physicians (42.9%) compared with lower rates in surgery (20.0%) and other specialties (23.1%).

When survey respondents were asked to select challenges they face in current use of AI tools, "information inaccuracy" was the most common concern overall, cited by 71.4% of IM physicians, 60.0% of surgeons, and 69.2% of other specialists. "Medicolegal implications" and "bias in responses" were reported most frequently by IM (50.0%) and surgery (60.0%), while "patient privacy" was cited far more often by "Other" specialties (53.8%) compared with IM (21.4%) and surgery (10.0%). "Lack of training" and "integration issues" were reported at comparable levels across all groups, though both were slightly higher in the "Other" category (46.2%). These statistics are as seen in Figure 2.

3.2 Qualitative Findings: Thematic Analysis of Specialty-Specific AI Applications

Of the 52 participants, 36 provided responses to the open-ended question, "What are some unique aspects of AI that you are using in your current specialty?" After manual categorization by two researchers, further analysis augmented by Google Gemini yielded four major themes: (1) AI as a Documentation & Communication Assistant; (2) Knowledge Synthesis & Decision Support; (3) Workflow Optimization; and (4) Barriers & Mistrust (Figure 3).

Theme 1 - AI as a Documentation & Communication Assistant: This theme was noticed across specialties, highlighting the use of LLMs to manage multiple facets of clinical work. A neurology resident noted using AI for "discharge summaries and patient education," adding that while the output

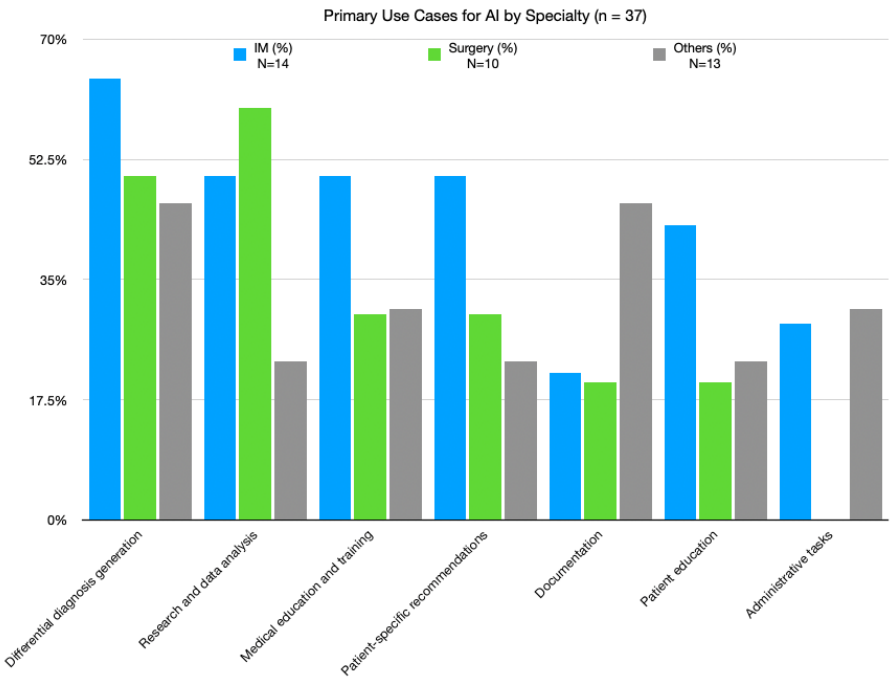

Figure 1: Current Clinical Applications of AI Tools Reported by Respondents (N=37), stratified by medical specialty. Survey respondents were asked to select all cases for which they currently use AI tools.

is “never 100% accurate,” it "significantly cut down on my time documenting.” Another respondent in otolaryngology reported “using ambient AI for note generation.” Another user noted LLMs to be used for creating “great discharge patient instructions.”

Theme 2 - Knowledge Synthesis & Decision Support: This theme captures the use of LLMs as an interactive tool for clinical reasoning. An IM resident described using ChatGPT and Open Evidence to help with "differential building and diagnostic testing," as well as “searching for evidence of drug-drug interactions with lesser known medications.” Another clinician used it for “summarization of guidelines [and] validation of differential diagnoses.” This application was particularly noted for complex or unusual cases where clinicians sought to broaden their considerations. However, respondents consistently indicated that these tools were used for idea generation rather than final decision-making.

Theme 3 - Workflow Optimization: This theme illustrates the application of LLMs in tasks that are ancillary to direct patient care. A significant number of participants reported using these tools to address administrative burdens. A pediatrician found that “using ChatGPT to write medical letters of necessity has been time-saving,” a sentiment echoed by an IM physician who used it to “help with prior authorizations and appeal letters.” Other uses included academic and research tasks, such as “literature review and writing analytic code in R and Python.”

Theme 4 - Barriers and Mistrust: The final theme reflects the tension between the perceived potential of AI and the significant barriers to its use. Some participants expressed strong mistrust and ethical opposition. One pediatrician stated simply, “I don’t use AI at work because I don’t trust it.” This view was articulated by an IM resident who believes AI use is “both unethical and dangerous to our field and the environment.”

4 Discussion

This mixed-methods study offers detailed insight into clinician perspectives on LLMs, revealing a pattern of pragmatic adoption limited by significant concerns about safety, reliability, and oversight. Our sample of 52 physicians represented a diverse range of specialties and practice settings, with

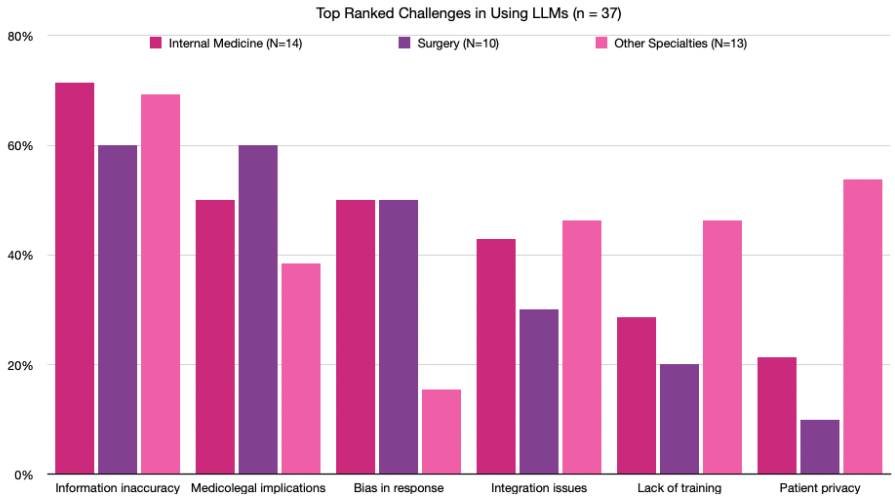

Figure 2: Current clinical challenges of AI tools reported by respondents (N=37), stratified by medical specialty. Survey respondents were asked to indicate the challenges or concerns they encounter in their current use of AI tools.

IM and surgery as the largest subgroups. A majority (71.2%) reported using at least one AI tool in clinical practice, with ChatGPT as the most familiar and frequently adopted. The findings suggest that clinicians are beginning to integrate these tools into their workflows. However, this adoption is driven more by the potential for efficiency gains in administrative tasks, rather than a deep-seated trust in the technology for high-stakes clinical decision-making.

Understandably, the deployment of these technologies in clinical settings presents substantial challenges. A primary concern is the factual accuracy of LLM-generated content, as these models are known to produce hallucinations [8]. Additionally, the lack of transparency in AI models raises concerns regarding clinician trust in LLMs and accountability in clinical decision-making [9]. This contributes to significant medicolegal uncertainty, as existing liability frameworks are not well-equipped to address errors originating from AI systems [10]. Furthermore, the potential for algorithmic bias to perpetuate health disparities and risks to patient data privacy under regulations like the Health Insurance Portability and Accountability Act represents a critical barrier to safe and equitable implementation [11].

A central finding is the dichotomy between clinicians' use of LLMs and their stated lack of trust. The quantitative data show that a majority of users engage with these tools for tasks like assistance with differential diagnoses and documentation, which are directly tied to the administrative burdens associated with burnout [12]. Yet, the most frequently cited challenges were information inaccuracy and medicolegal implications, highlighting the tension between workload relief and the foundational requirements of clinical reliability. The qualitative data support this interpretation. One participant noted that AI-generated summaries "are never 100% accurate" but "significantly cut down on...time documenting," explicitly stating that the output requires editing. This behavior reflects a "human-in-the-loop" model, where the clinician acts as the final validator of AI-generated content. Our findings confirm that clinicians operate under this assumption, critically appraising AI outputs. This aligns with recent clinician surveys showing that LLMs are viewed primarily as assistive tools for cognitive support and education rather than autonomous systems; Spotnitz et al. (2024) and Sumner et al. (2025) similarly report enthusiasm for the application of this technology, tempered by caution surrounding its implementation and the ongoing need for human-centered development [13, 14]. However, our data also expose a potential vulnerability in the implementation of this "human-in-the-loop" model. A lack of training was identified as a key challenge by nearly a third of AI users in our survey, a barrier similarly well-documented in the literature [5]. While the "human-in-the-loop" model remains the widely endorsed framework for medical AI integration, our findings suggest that its success is limited

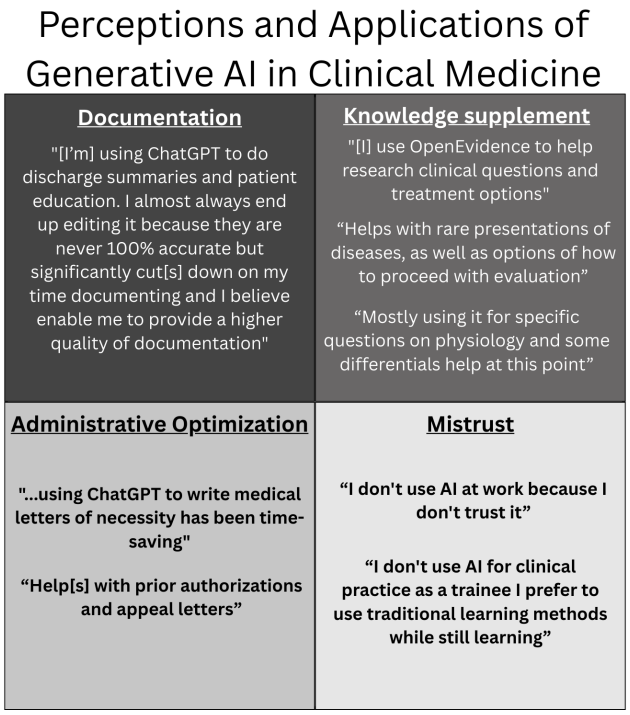

Figure 3: Identified Themes of AI & LLMs, with quotations from different survey respondents.

by inadequate training and support, underscoring a critical gap between conceptual endorsement and practical implementation.

Adoption patterns varied across specialties, although not to a statistically significant degree. It is important to acknowledge the overrepresentation of IM and surgical specialties, thus our findings may not fully capture perspectives from fields underrepresented in our study, such as pediatrics or psychiatry. Regardless, our data reveal that AI adoption is not a monolithic phenomenon; rather, usage patterns appear to diverge based on the specific demands of different clinical specialties. Clinicians in IM, for example, reported the highest rate of using LLMs for cognitive support tasks such as "differential diagnosis generation" (64.3% of respondents) and "medical education and training" (50.0%). This may be attributable to the nature of the specialty, which often involves managing diagnostically complex cases with undifferentiated symptoms, requiring the synthesis of a vast and rapidly evolving evidence base. In this context, LLMs may serve as a "digital consultant" to broaden clinical reasoning or quickly access recent guidelines.

While these differences are notable, it is equally important to highlight the relative uniformity across specialties. The overall magnitude of variation was small, and this likely reflects the fact that a large proportion of our respondents were in residency or fellowship training. Trainees across disciplines often share common workflow needs, such as support with differential diagnosis generation and management plans, while being less involved in administrative responsibilities that are more prominent later in practice. As such, the training-heavy composition of our sample may have shifted responses toward cognitive support tasks across the board.

4.1 Strengths and Limitations

This study has several strengths. Its mixed-methods design allowed for both quantitative and qualitative exploration of physician engagement with large language models, offering a multi-dimensional view of adoption patterns, perceived utility, and barriers. The inclusion of participants from a broad range of medical specialties adds further depth, allowing us to capture specialty-specific applications and concerns that may not be evident in single-specialty studies.

Nonetheless, several limitations should be acknowledged. First, the use of a non-probability, convenience sampling strategy limits the generalizability of our findings. Our sample may be skewed toward individuals with a pre-existing interest in AI, as physicians who already use or are curious about these tools may be more likely to respond. Relatedly, participants who did not want to disclose the use of AI in their specialty sometimes declined to complete the survey, introducing a non-response bias that could underrepresent more skeptical perspectives. Second, all data were self-reported and therefore subject to recall bias and social desirability bias, particularly in reporting attitudes toward AI use in patient-facing contexts. Third, while we included respondents across multiple specialties, the sample sizes within each specialty were modest. Several specialties were represented by only a single respondent, limiting the depth of interpretation for those fields.

Given the overrepresentation of IM and surgical specialties, our findings may not fully capture perspectives from underrepresented fields such as pediatrics or psychiatry. Additionally, we recognize that using an LLM to analyze qualitative perceptions of LLMs is not a widely validated strategy and may introduce or perpetuate bias, as model outputs can reflect the same linguistic or conceptual biases inherent to the underlying training data. To mitigate this, all generated themes were reviewed for alignment with participant language and intent, and discrepancies were discussed until consensus was achieved. Lastly, as a cross-sectional study, our findings reflect perceptions at a single point in time; given the rapid evolution of AI technologies, clinician perspectives and usage patterns are likely to shift quickly, limiting the temporal stability of these results. While our results highlight emerging themes around clinician engagement with LLMs, they should be interpreted within the context of our sample and may differ in other specialties or health systems.

To address these limitations, further research can expand the scope and methodological rigor of this work. For instance, future studies should aim to recruit larger and more balanced specialty cohorts; this will allow for more robust specialty-specific insights, as well as more representative comparisons and stronger inferences about differences in adoption. Additionally, it would be beneficial to include a larger cohort of attending physicians, which may reveal a broader distribution of use cases that more accurately reflects the full spectrum of specialty-specific workflows. Lastly, we could expand upon our qualitative data analysis with standardized qualitative analysis frameworks, such as NVivo, in order to enhance transparency and reproducibility.

5 Conclusion

Clinicians are engaging with LLMs in a resourceful, but cautious manner. They are pragmatically adopting these tools to alleviate administrative burdens, particularly in documentation and knowledge synthesis. This adoption, however, exists in tension with a profound and justified mistrust regarding the technology's accuracy, the lack of formal training for its use, and an unclear medicolegal landscape. The prevailing safety model, which relies on a "human-in-the-loop," is compromised by the absence of adequate training and support for this critical oversight role. Moving beyond the current stage of cautious adoption toward safe and effective integration will require a collaborative effort from developers, healthcare organizations, educators, and policymakers to build trustworthy tools and the frameworks necessary to support their responsible use in clinical practice.

Acknowledgments and Disclosure of Funding

This work was made possible through collaborations including H2AI at Georgetown University, an AI and medicine platform fostering innovation, research, and education.

References

[1] American Medical Association. (2025). *Physician sentiments around the use of AI in heath care: motivations, opportunities, risks, and use cases*. American Medical Association.

[2] Large Language Models Applied to Health Care Tasks May Improve Clinical Efficiency, Value of Care Rendered, Research, and Medical Education. (2025). *PubMed*. Retrieved August 23, 2025, from https://pubmed.ncbi.nlm.nih.gov/39694303/

[3] Artificial Intelligence (AI) - Powered Documentation Systems in Healthcare: A Systematic Review. (2025). *PubMed*. Retrieved August 23, 2025, from https://pubmed.ncbi.nlm.nih.gov/39966286/

[4] 25 × 5 Symposium to Reduce Documentation Burden: Report-out and Call for Action. (2022). *PMC*. Retrieved August 23, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC9095342/

[5] Overcoming barriers and enabling artificial intelligence adoption in allied health clinical practice: A qualitative study. (2025). *PMC*. Retrieved August 23, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC11792011/

[6] Wadhwa V, Alagappan M, Gonzalez A, Gupta K, Brown JRG, Cohen J, Sawhney M, Pleskow D, Berzin TM. (2020). Physician sentiment toward artificial intelligence (AI) in colonoscopic practice: a survey of US gastroenterologists. *Endoscopy International Open*, 8(10), E1379-E1384.

[7] Buck C, Doctor E, Hennrich J, Jöhnk J, Eymann T. (2022). General Practitioners' Attitudes Toward Artificial Intelligence-Enabled Systems: Interview Study. *Journal of Medical Internet Research*, 24(1), e28916.

[8] Evaluating Large Language Models on Medical Evidence Summarization. (2023). *PMC*. Retrieved August 23, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC10168498/

[9] Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. (2025). *PubMed Central*. Retrieved August 23, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC11977975/

[10] Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. (2023). *PMC*. Retrieved August 23, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC10711067/

[11] Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. (2025). *PMC*. Retrieved August 23, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC12076083/

[12] Burnout Related to Electronic Health Record Use in Primary Care. (2023). *PMC*. Retrieved August 23, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC10134123/

[13] A Survey of Clinicians' Views of the Utility of Large Language Models (2024). *PMC*. Retrieved October 01, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC11023712/

[14] Perspectives and Experiences With Large Language Models in Health Care: Survey Study. (2025). *PMC*. Retrieved October 01, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC12082058/

A Survey Instrument

Section I: Demographics

What is your professional degree? MD/DO, NP/PA/APP, Other

What is your primary specialty? Allergy and Immunology, Anesthesiology, Dermatology, Diagnostic Radiology, Emergency Medicine, Family Medicine, Internal Medicine, Medical Genetics, Neurology, Nuclear Medicine, Obstetrics and Gynecology, Ophthalmology, Pathology, Pediatrics, Physical Medicine and Rehabilitation, Preventive Medicine, Psychiatry, Radiation Oncology, Surgery, Urology, Other

How long have you been practicing in your specialty? Training stage (residency or fellowship), 0-5 years, 6-10 years, 11-20 years, 20+ years

A rough estimate of your clinical/research time split: 0% clinical / 100% research, 25% clinical / 75% research, 50% clinical / 50% research, 75% clinical / 25% research, 100% clinical / 0% research

A rough estimate of your inpatient/outpatient time split: 0% outpatient / 100% inpatient, 25% outpatient / 75% inpatient, 50% outpatient / 50% inpatient, 75% outpatient / 25% inpatient, 100% outpatient / 0% inpatient

Section II: Experience with Artificial Intelligence Tools

How familiar are you with the following tools: *(Options: Not Familiar, Somewhat Familiar, Very Familiar for each of the following)*

- ChatGPT
- Perplexity
- Claude
- Microsoft Copilot

Please select the tool you find most helpful to you, as the following questions will apply to it. *If you are not familiar with any tool, select "None." If you primarily use another tool, you can write the name in "Other."*
ChatGPT, Perplexity, Claude, Microsoft Copilot, None, Other

What do you use the selected tool for in clinical practice? *(Select all that apply)*
Documentation, Differential diagnosis, Patient-specific recommendations, Research and data analysis, Medical education and training, Patient education, Communication with colleagues, Administrative tasks, Other, N/A

How often do you use the tool selected in your clinical practice? Daily, Several times a week, Once a week, Occasionally, Never, N/A

What challenges do you see with using the selected tool in your practice? *(Select your top three)*
Information inaccuracy (lack of citations), Medicolegal implications, Lack of training, Patient privacy, Bias in response, Integration issues, Other, N/A

How do you document your use of AI? Orally, Written (i.e. in EMR), Neither, N/A

To whom do you communicate your use of AI? Patients, Physicians and other colleagues, N/A

What features are most important to you in an AI/LLM? *(Select all that apply)*
Documentation, Differential diagnosis, Patient-specific recommendations, Research and data analysis, Medical education and training, Patient education, Communication with colleagues, Administrative tasks, Other

Section III: Specialty Specific Questions

What are some unique aspects of AI that you are using in your current specialty? *Feel free to use this space to expand on prior drop-down answers or provide new answers.*

B Supplemental Tables

Table 1: Participant Demographic and Professional Characteristics

Characteristic	Category	N (%)
Professional Degree	MD/DO	52 (100)
Primary Specialty (N=52)	Internal Medicine	17 (32.7)
	Surgery	14 (26.9)
	Pediatrics	9 (17.3)
	Emergency Medicine	4 (7.7)
	Other	8 (15.4)
Years in Practice (N=47)	Training stage	25 (53.2)
	0-5 years	7 (14.9)
	6-10 years	5 (10.6)
	11-20 years	7 (14.9)
	20+ years	3 (6.3)
Clinical/Research Split (N=48)	100% clinical	31 (64.5)
	75% clinical / 25% research	12 (25.0)
	50% clinical / 50% research	2 (4.2)
	25% clinical / 75% research	2 (4.2)
	100% research	1 (2.1)
Inpatient/Outpatient Split (N=52)	100% outpatient	9 (17.3)
	75% outpatient / 25% inpatient	11 (21.2)
	50% outpatient / 50% inpatient	4 (7.7)
	25% outpatient / 75% inpatient	17 (32.7)
	100% inpatient	11 (21.2)

Table 2: Familiarity with Specific LLMs (N=52)

	Not Familiar (%)	Somewhat Familiar (%)	Very Familiar (%)
ChatGPT	5 (9.6)	10 (19.2)	7 (13.5)
Perplexity	42 (80.8)	2 (3.8)	4 (7.7)
Claude	46 (88.5)	3 (5.8)	0 (0)
Microsoft Copilot	30 (57.7)	15 (28.8)	4 (7.7)

Table 3: Frequency of Use of Primary AI Tool (N=37)

Frequency	N (%)
Daily	10 (27.0)
Several times a week	4 (10.8)
Once a week	5 (13.5)
Occasionally	11 (29.7)
Never	7 (18.9)

Table 4: P-values comparing use cases across specialties (N=37)

Use Case	P-value (IM vs Surgery)	P-value (IM vs Other)	P-value (Surg vs Other)
Differential diagnosis	0.678	0.449	1
Research and data analysis	0.697	0.236	0.102
Medical education and training	0.421	0.440	1
Patient-specific recommendations	0.421	0.236	1
Documentation	1	0.236	0.379
Patient education	0.388	0.420	1
Administrative tasks	0.114	1	0.104

Table 5: Most Important AI/LLM Features as Rated by Clinicians (N=52)

Feature	Internal Med N=17	Surgery N=13	Other N=22	Total N=52
Medical education & training	58.8%	46.2%	40.9%	48.1%
Administrative tasks	58.8%	53.8%	31.8%	46.2%
Differential diagnosis	52.9%	30.8%	45.5%	46.2%
Patient education	52.9%	15.4%	59.1%	44.2%
Documentation	47.1%	38.5%	45.5%	44.2%
Research and data analysis	41.2%	61.5%	27.3%	40.4%
Patient-specific recommendations	41.2%	23.1%	36.4%	34.6%
Communication with colleagues	17.6%	15.4%	13.6%	15.4%

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's contributions are related to examining clinician impressions of artificial intelligence (namely, LLMs) and this has been captured accurately in the abstract/introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Strengths and Limitations section

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theoretical results

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Survey provided in appendix

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Survey and data provided in appendix/results

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: No training/test data

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: P-values are reported for differences between IM/Surg

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: No compute

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All ethical guidelines were followed, and all participants were willing and able participants. IRB approval was sought and obtained prior to dissemination of the survey.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The discussion offers both positive and negative viewpoints of clinicians on the use of AI/LLMs in medicine.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No concerns for high-risk data/models

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All creators of the survey are credited in authorship.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Methods section

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: IRB approval was sought and obtained prior to dissemination of the survey. No risks were associated with taking the survey, and data was never identifiable.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The Methods section explicitly states that Google Gemini 2.5 Pro was used for thematic analysis of qualitative free-text responses, which is a part of our mixed-methods analysis.