How are doctors across specialties using commercial large language models? Insights from the Anthropic Economic Index

Izabella Mancewicz
Monash University
Yufei Xu
Monash University
Jeff R. Ma
Monash University
Khoa N. Cao
khoa.cao@monash.edu
Monash University

Research Article

Keywords:

Posted Date: March 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-7384730/v1

Additional Declarations: No competing interests reported.

Abstract

Background

Commercial large language models (LLMs) have demonstrated potential across a range of medical applications, yet empirical data on their real-world clinical use remains limited. Understanding the extent and nature of LLM adoption by doctors across medical specialties could reveal practical insights into the implementation of AI in healthcare.

Methods

Using data from the Anthropic Economic Index (March 2025 release), we analyzed approximately one million anonymized conversations with Claude 3.7 from February 2025, categorizing interactions by medical specialty, task intent, and type of AI interaction (augmentation vs. automation). Medical specialties and tasks were aligned with the Occupational Information Network (O*NET) database and adjusted for workforce size.

Results

Out of one million conversations, 5,998 (0.60%) were attributed to medical doctors, covering 17 specialties and 65 distinct clinical tasks. The majority of conversations involved diagnostics/interpretation (41.1%) and communication/consultation (27.6%), contrary to existing literature emphasizing administrative uses. Specialties such as radiology, allergology/immunology, and pathology showed the highest absolute usage of Claude, while allergists/immunologists, pathologists and nuclear medicine physicians had the highest utilization when adjusted for workforce size. Physicians predominantly used Claude to learn, validate, or perform iterative tasks rather than to automate tasks, with significant variation by specialty.

Conclusion

Our findings indicate predominant use of a commercial LLM in clinical tasks, particularly diagnostics and patient communication, rather than in administrative tasks only. This study highlights differential adoption rates among specialties and patterns of augmentation compared to automation, marking critical areas for future AI integration and development.

Introduction

Significant advances in medical artificial intelligence (AI) have emerged over the past few years, particularly with the rise of large language models (LLMs) trained on the massive corpora of publicly available text. LLMs have been explored in a wide range of clinical tasks, ranging from documentation[1] responding to patient queries[2],to making diagnoses from patient history[3]. Despite these advances, the extent and nature of commercial LLM use by the medical field is largely unknown beyond anecdotal evidence.

Claude is a family of widely used large language models (LLMs) developed by the AI research company Anthropic. They are designed to engage in natural language understanding and generation, perform tasks such as question answering, text creation and summarization, coding support, and reasoning with large

context windows.[10] The recent release of the Anthropic Economic Index[4,5] provides data on the usage of Claude in occupations across the labor market. Leveraging this dataset, this study aims to 1) quantify the Claude interactions attributed to tasks performed by different medical specialties, 2) analyze the type of clinical tasks (intent domain) represented in Claude interactions, and 3) describe the type of user interaction (automation vs augmentation) demonstrated in these conversations. While the dataset captures interactions with one LLM in a limited time window, the results provide novel insights into the real-world usage of Claude in medical tasks and indicates possible trends in LLM usage in medicine.

Methods

Data Sources

Data was sourced from Anthropic Economic Index,[4,5] an open-source dataset released by Anthropic, Inc. (San Francisco, California). Data was drawn primarily from the March 27th 2025 release,[6] which examined approximately one million anonymized Claude.ai Free and Pro conversations following the release of Claude 3.7 in February 10th 2025 and reported on the percentage of conversations attributed to a specific occupation and work-related task as categorized by the Occupational Information Network (O*NET), a database maintained by the US Department of Labor. Data was also drawn from the February 2025 release (covering December - January 2024)[7] to examine trends over time. The Anthropic Economic Index dataset provides pre-mapped data where conversations with Claude have been mapped onto O*NET occupational tasks by Anthropic using Clio, a privacy-preserving model that extracts characteristics in natural language conversations.[11] Each conversation, if occupationally relevant, is mapped to one task. Tasks mapped to less than 15 conversations or less than 5 user accounts were excluded for privacy. Conversations were labelled by Anthropic by user interaction type, specifically whether the interaction represented “automation” or “augmentation”[6]. Automation in this dataset refers to where AI directly complete tasks with minimal human involvement, through a “directive” or a “feedback loop”; while augmentation refers to where AI enhances the human user’s work through a collaborative process, through “task iteration”, “learning”, and “validation”.[8] The accuracy of task-mapping and user interaction classification was 86% and 90.7%, respectively.[8] The number of specialists in each medical specialty were sourced from the Bureau of Labor of Statistics (BLS)[9] and where absent, other relevant sources as detailed in Supplement S1.

Pre-processing

Out of 65 medical tasks identified in the March data release, 58 tasks were mapped one-to-one onto a specialty, and 7 were mapped onto more than one occupation. The total percentage of conversations attributed to these tasks were split among these multiple occupations based on the size of these specialties. The percentage of conversations attributable to each specialty among the total percentage of conversations this task accounts for were calculated with the following formula:

$$\text{Adjusted conversation percentage of a specialty} = \frac{original\ percentage}{workforce\ sizes\ of\ all\ specialties\ sharing\ this\ task} \times size\ of\ this\ specialty$$

A worked example of this process can be found in Supplement S3b.

In addition to the five user-interaction types under which tasks were categorized in the dataset by Anthropic, this study categorizes tasks under five higher-order tasks groups called intent domains. The five higher-order task groups were derived inductively by clustering tasks that exhibited recurring patterns and shared characteristics. This thematic grouping organized individual tasks into broader categories reflecting commonalities in purpose, structure, or cognitive demands. Each task was subsequently independently categorized by two authors (IM, YX) into one of five intent domains, with an inter-rater reliability of 85.9%. Discrepancies were resolved with discussion with a third author (KC). The five intent domains include: *communication/consultation, clinical management/treatment planning, diagnostics/interpretation, research/education* and *administrative/reporting*. Tasks were attributed by identifying keywords such as “develop treatment plans,” “coordinate care” for *clinical management/treatment planning* and “analyze” and “diagnostic results” for *diagnostics/interpretation*.

Data analysis

We completed several analyses. First, the most common tasks were identified across all occupations and on a per-occupation basis. Second, Claude usage by specialty was calculated by summing the number of conversations for every task completed by an occupation. The number of conversations for each task was determined using percentage of all conversations the task accounts for, which is provided in the original dataset, and multiplying this by the total number of conversations of 1 million. The number of conversations per 100,000 specialists was also calculated. The total number of conversations were subsequently adjusted by the specialty size, with a ‘normalized ratio’ whereby the specialty occupation with the lowest use of Claude was assigned a reference value of 1.0. The normalized ratio is calculated using adjusted AI usage of a specialty, divided by adjusted AI usage of the reference specialty (Supplement S4). Task coverage was also determined by dividing the quantity of tasks reported in the Anthropic Economic Index by the total quantity of tasks reported in the O*NET data to determine which occupation utilized Claude across the greatest proportion of their day-to-day tasks. Third, tasks were assessed across the five domains of intent and the top five specialties were identified for each domain. Fourth, medical tasks were assessed by type of user interaction overall and per-occupation. Finally, concordance and qualitative differences in Claude usage between February and March data were reported, followed by an assessment of Claude usage in doctors compared to wider society. Analyses were performed on data from both February and March releases for Claude usage by tasks, specialty, and intent domains. Analysis of user interaction was only performed on March release data due to user interaction type not being reported in the February release. Analyses were performed in Microsoft Excel for Mac version 16.102.2.

Results

From the 1 million conversations released in the Anthropic Economic Index, 0.60% (n=5,998) were attributed to tasks completed by medical doctors. Conversations were reported for 17 out of 21 (81.0%) medical specialties and 65 out of 322 (20.2%) medical tasks from the O*NET database. The number of conversations for individual medical tasks ranged from 15 (least common) to 652 (most common). The most common medical tasks overall are reported in Table 1 and on a per-occupation basis in Supplement S2. Percentages refers to share of conversations on this task among all medical conversations.

Table 1. Most common medical tasks performed by Claude

	Task	Occupation/Medical Specialty	Intent Domain	No. of Conversations (%)
1	Prepare comprehensive interpretive reports of findings.	Radiologists; Nuclear Medicine Physicians	Diagnostics/Interpretation	652 (10.9%)
2	Educate patients about diagnoses, prognoses, or treatments.	Allergists and Immunologists	Communication/Consultation	441 (7.4%)
3	Analyze and interpret results from tests such as microbial or parasite tests, urine analyses, hormonal assays, fine needle aspirations (FNAs), and polymerase chain reactions (PCRs).	Pathologists	Diagnostics/Interpretation	387 (6.5%)
4	Interpret diagnostic test results to make appropriate differential diagnoses.	Allergists and Immunologists	Diagnostics/Interpretation	365 (6.1%)
5	Educate patients about maintenance and promotion of healthy vision.	Ophthalmologists	Communication/Consultation	276 (4.6%)
6	Teach, take continuing education classes, attend conferences or seminars, or conduct research and publish findings to increase understanding of mental, emotional, or behavioral states or disorders.	Psychiatrists	Research/Education	252 (4.2%)
7	Analyze records, reports, test results, or examination information to diagnose medical condition of patient.	Internists; Obstetricians and Gynecologists	Diagnostics/Interpretation	246 (4.1%)
8	Communicate examination results or diagnostic	Radiologists	Communication/Consultation	235 (3.9%)

	information to referring physicians, patients, or families.			
9	Examine microscopic samples to identify diseases or other abnormalities.	Pathologists	Diagnostics/Interpretation	205 (3.4%)
10	Develop individualized treatment plans for patients, considering patient preferences, clinical data, or the risks and benefits of therapies.	Allergists and Immunologists	Clinical Management/Treatment Planning	195 (3.3%)

In absolute terms, radiologists (n = 1138, 19.0% of all medical conversations), allergists/immunologists (n = 1067, 17.8%) and pathologists (n = 818, 13.6%) were the specialties most likely to utilize Claude (Table 2). With adjustment for specialty size, allergists/immunologists, pathologists and nuclear medicine physicians were respectively 236.0, 116.5 and 105.6 times more likely to utilize Claude than the specialty with the lowest recorded use of AI (hospitalists). General pediatricians and family/general practitioners utilized Claude across the greatest proportion of their day-to-day tasks. No conversations were attributed to cardiologists, dermatologists, emergency medicine physicians or pediatric surgeons.

Table 2: Claude usage ranking by specialty

Occupation/Medical Specialty	No. of Conversations (%)	No. of Conversations per 100,000 Specialists	Normalized Ratio*	O*NET Task Coverage (%)
Radiologists	1138 (19.0%)	3559	55.9	11/29 (37.9%)
Allergists and Immunologists	1067 (17.8%)	15030	236.0	4/16 (25%)
Pathologists	818 (13.6%)	7419	116.5	7/19 (36.8%)
Ophthalmologists	519 (8.7%)	4498	70.6	5/18 (27.8%)
Psychiatrists	381 (6.4%)	1535	24.1	3/12 (25%)
Family and General Practitioners	375 (6.3%)	321	5.0	7/15 (46.7%)
Internists, General	353 (5.9%)	511	8.0	7/19 (36.8%)
Pediatricians, General	334 (5.6%)	900	14.1	9/16 (56.3%)
Nuclear Medicine Physicians	251 (4.2%)	6725	105.6	3/26 (11.5%)
Obstetricians and Gynecologists	185 (3.1%)	641	10.1	6/15 (40%)
Preventive Medicine Physicians	165 (2.8%)	1701	26.7	4/15 (26.7%)
Physical Medicine and Rehabilitation Physicians	115 (1.9%)	1078	16.9	2/15 (13.3%)
Neurologists	91 (1.5%)	978	15.4	3/24 (12.5%)
Surgeons	41 (0.7%)	76	1.2	2/13 (15.4%)
Anesthesiologists	32 (0.5%)	95	1.5	1/18 (5.6%)
Hospitalists	32 (0.5%)	64	1.0	1/13 (7.7%)
Urologist	26 (0.4%)	188	3.0	1/14 (7.1%)
Sports Medicine Physicians	24 (0.4%)	648	10.2	1/27 (3.7%)

*AI usage of specialties was normalized with reference to the adjusted AI usage of hospitalists, which is adjusted by the size of this specialty. This was due to hospitalists having the lowest number of

conversations per 100,000 specialists. The workforce size of the hospitalist specialty was derived from an external report, see Supplement S1.

The intent of conversations was distributed across diagnostics/interpretation (41.1%), communication/consultation (27.6%), clinical management/treatment planning (15.2%), administrative/reporting (10.8%) and research/education (5.4%). The top specialties in each domain of intent are reported in Table 3.

Table 3: Top specialties for each intent domain

	Diagnostics/ Interpretation	Communication/ Consultation	Clinical Management/ Treatment Planning	Administrative/ Reporting	Research/ Education
1	Pathologists	Allergists and Immunologists	Allergists and Immunologists	Allergists and Immunologists	Psychiatrists
2	Allergists and Immunologists	Ophthalmologists	Pediatricians, General	Preventive Medicine Physicians	Pathologists
3	Nuclear Medicine Physicians	Nuclear Medicine Physicians	Ophthalmologists	Physical Medicine and Rehabilitation Physicians	
4	Radiologists	Radiologists	Obstetricians and Gynecologists	Radiologists	
5	Ophthalmologists	Preventive Medicine Physicians	Preventive Medicine Physicians	Urologists	

71.9% of conversations were categorized as augmentation, while 28.1% of conversations were categorized as automation. The majority of user interactions fell under learning (augmentation, 51.9% of all conversations), directive (automation, 27.4%) and task iteration (augmentation 18.0%). Type of user interaction by medical specialty is detailed in Figure 1. Pathologists, physical medicine and rehabilitation physicians, nuclear medicine physicians and radiologists were most likely to automate their tasks. In comparison, hospitalists, neurologists, preventative medicine physicians and sports medicine physicians demonstrated no automation in included tasks, noting that some tasks were excluded from the user-interaction analysis in the original dataset.

The findings were relatively concordant between the February and March datasets in terms of specialty representation. For Claude usage by specialty, the 10 specialties with the highest usage in February remained the same in March with some ranking changes (i.e. a rise for ophthalmology and psychiatry, and a fall for family/general practitioners and general internists). Between the February and March datasets, the percentage of conversations attributed to physicians increased from 0.52% to 0.60%, with a rise in

coverage from 54 to 65 tasks. When compared to the general population, doctors represented 0.26% of the American workforce but accounted for 0.60% of conversations with Claude. This indicated proportionally greater use of Claude by doctors compared to other healthcare workers, including medical record and health information technicians, office workers and management personnel but lower than computer science and mathematics (3.4% of US workforce, 37.2% of Claude usage), and life, physical and social sciences (0.9% of US workforce, 6.4% of Claude usage). Doctors were also automating fewer tasks than some other professionals, with only 28% of conversations automated compared to 43% in the entire workforce.

Discussion

As the Anthropic Economic Index only includes data on usage of Claude, these findings do not capture the use of other LLMs, such as ChatGPT or open-source alternatives, in medical tasks. The patterns and trends identified in this study may be representative of doctors' overall usage of LLMs, given Claude is one of the most used LLM models with an estimate of 18.9 million monthly active users.[12] However, as conversations are anonymized in this dataset, demographic information on users engaging in medically related conversations are not available. This makes it challenging to determine with certainty the representativeness of this cohort compared to doctors who use other LLMs. Nevertheless, these findings based on data from Claude signal a potential shift in how doctors are interacting with commercial LLMs. While many prior publications such as the 2025 AMA Physician Sentiment Survey[13] and 2024 HIMSS AI Adoption in Healthcare Report[14] have emphasized the use of AI primarily for administrative tasks such as billing, transcription, and documentation, the Anthropic dataset reveals higher-than-expected utilization of an LLM for diagnosis, interpretation and communication. A significantly greater proportion of conversations were categorized under clinical tasks than administrative tasks. This could indicate potential utility and functionality of LLMs in all aspects of medicine, including clinical communication, diagnosis and treatment planning.

To the authors' knowledge, no prior studies have systematically examined the real-world usage pattern of any LLM in medicine across disciplines. Previous analyses have predominantly focused on AI research output[15] or regulatory authorizations (e.g. FDA authorization[16]) as a surrogate measure of AI adoption in healthcare. However, such proxies are unlikely to accurately represent how clinicians are currently engaging with commercial LLMs in practice. Allergists/immunologists, pathologists, ophthalmologists and psychiatrists ranked amongst the medical specialties with the highest utilization of Claude in this study, despite relatively few FDA-authorized products on the market or publications within the medical AI corpus. This discrepancy suggests that real-world adoption is likely de-coupled from formal regulatory approval or academic research activity, indicating a potential lag in existing frameworks to capture or govern the full spectrum of AI use in medicine. Claude task coverage was highest for general pediatricians and family/general practitioners, suggesting that commercial models may already offer practical utility across a wide array of clinical scenarios, particularly in high-volume settings wherein efficiency is highly valued. In contrast, surgery and procedural specialty such as anesthesiology reported the lowest use of Claude. This pattern aligns with the current limitations of LLMs, which primarily excels in cognitive,

language-based tasks rather than interventional domains. Whether this gap will narrow in the future with progress in embodied AI or advances in robotic integration remains an open question.

Compared to professions in other sectors such as computer science and finances, medical tasks exhibited lower levels of task automation, suggesting that Claude is currently predominantly used to augment rather than replace clinical work. Specialties demonstrating strong Claude usage overall, such as pathology and radiology, also display markedly higher automation use. One possible explanation for this is the high degree of digitization, structured workflows, and visual data formats in these specialties align well with existing AI capabilities.

This study has several limitations. First, attribution of conversations to specific occupations was based on task-matching algorithms and was anonymized to preserve user privacy. The exact user identity and clinical scenario is unknown, risking mis-classification. Second, only conversations from Claude Free and Pro tiers were captured, rather than API, Team or Enterprise users. However, only conversations relevant to occupational tasks were filtered for the dataset, which may mitigate the first two limitations and filter out conversations not at a professional level. Third, taxonomy gaps remain as the ONET occupation list used by the Index lacks codes for several specialties (e.g. dermatologists, cardiologists, surgical subspecialties). Finally, privacy and frequency filters may have excluded sensitive or rare topics, potentially leading to under-representation of small specialties (e.g. mental health topics for psychiatrists). Integrating API traffic and expanding occupation taxonomies would provide a more complete picture of generative AI's penetration across medicine.

Conclusion

This paper provides one of the first specialty-level descriptive analyses of AI use in medicine based on real-world conversation data with Claude, a commercial large language model. By analyzing millions of anonymized chats released in the Anthropic Economic Index, our study reveals that the current use of a major commercial LLM in medicine extends beyond administrative tasks to diagnosis, treatment planning, and patient communication. Adoption varies widely across specialties, with allergists/immunologists, pathologists, and radiologists demonstrating highest usage. Claude is currently predominantly used to augment rather than automate clinical tasks. These insights provide a foundational understanding of real-world AI utilization in healthcare and highlight areas for further integration and development.

Declarations

Consent to Participate declaration: not applicable, as this study used published open-source data and did not collect new data.

Funding declaration: no funding

Human Ethics and Consent to Participate declarations: not applicable

Clinical trial number: not applicable

Availability of Data and Materials: The datasets analyzed in this study are publicly available under CC-BY licence from the Hugging Face Datasets repository, which are available at: https://huggingface.co/datasets/Anthropic/EconomicIndex

References

1. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Wilson Hannay SB, et al. Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses. NEJM Catalyst. 2025 Apr 16;6(5).

2. Small WR, Wiesenfeld B, Beatrix Brandfield-Harvey, Jonassen Z, Mandal S, Stevens ER, et al. Large Language Model–Based Responses to Patients' In-Basket Messages. JAMA network open. 2024 Jul 16;7(7):e2422399.

3. Tu T, Schaekermann M, Anil Palepu, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature. 2025 Apr 9.

4. Introducing the Anthropic Economic Index [Internet]. Anthropic. 2025 Feb 11. Available from: https://www.anthropic.com/news/the-anthropic-economic-index

5. Anthropic Economic Index: Insights from Claude 3.7 Sonnet [Internet]. Anthropic. 2025 . Available from: https://www.anthropic.com/news/anthropic-economic-index-insights-from-claude-sonnet-3-7

6. Anthropic. EconomicIndex Dataset, release 2025-03-27 [Internet]. Hugging Face; 2025 Mar 27. Available from:https://huggingface.co/datasets/Anthropic/EconomicIndex/tree/main/release_2025_03_27

7. Anthropic. Economic Index Dataset, release 2025-02-10 [Internet]. Hugging Face; 2025 Feb 10. Available from:https://huggingface.co/datasets/Anthropic/EconomicIndex/tree/main/release_2025_02_10

8. Handa K, Tamkin A, McCain M, Huang S, Durmus E, Heck S, et al. Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations [Internet]. arXiv.org. 2025. Available from: https://arxiv.org/abs/2503.04761

9. U.S. Bureau of Labor Statistics. May 2023 National Occupational Employment and Wage Estimates [Internet]. Bureau of Labor Statistics. 2023. Available from: https://www.bls.gov/oes/2023/may/oes_nat.htm

10. Anthropic's transparency hub [Internet]. Anthropic. 2025. Available from: https://www.anthropic.com/transparency/model-report

11. Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., Huang, S., Mountfield, A., Hong, J., Ritchie, S. and Stern, M., 2024. Clio: Privacy-preserving insights into real-world ai use. arXiv preprint arXiv:2412.13678.

12. Backlinko. Claude Statistics: How many people use Claude? [Internet]. Backlinko. 2025. Available from: https://backlinko.com/claude-users

13. AMA. AMA Augmented Intelligence Research, Physician sentiments around the use of AI in health care: motivations, opportunities, risks, and use cases [Internet]. American Medical Association. 2025 Feb. Available from: https://www.ama-assn.org/system/files/physician-ai-sentiment-report.pdf

14. HIMSS, Medscape. AI Adoption in Healthcare Report 2024 [Internet]. HIMSS; Medscape. 2024 Dec 6. Available from: https://cdn.sanity.io/files/sqo8bpt9/production/68216fa5d161adebceb50b7add5b496138a78cdb.pdf

15. Stewart JE, Rybicki FJ, Dwivedi G. Medical Specialties Involved in Artificial Intelligence Research: Is There a Leader. Tasman Medical Journal. 2020;2(1):20-27.

16. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Device [Internet]. FDA. 2025 March 25. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artific ial-intelligence-and-machine-learning-aiml-enabled-medical-devices

Figures

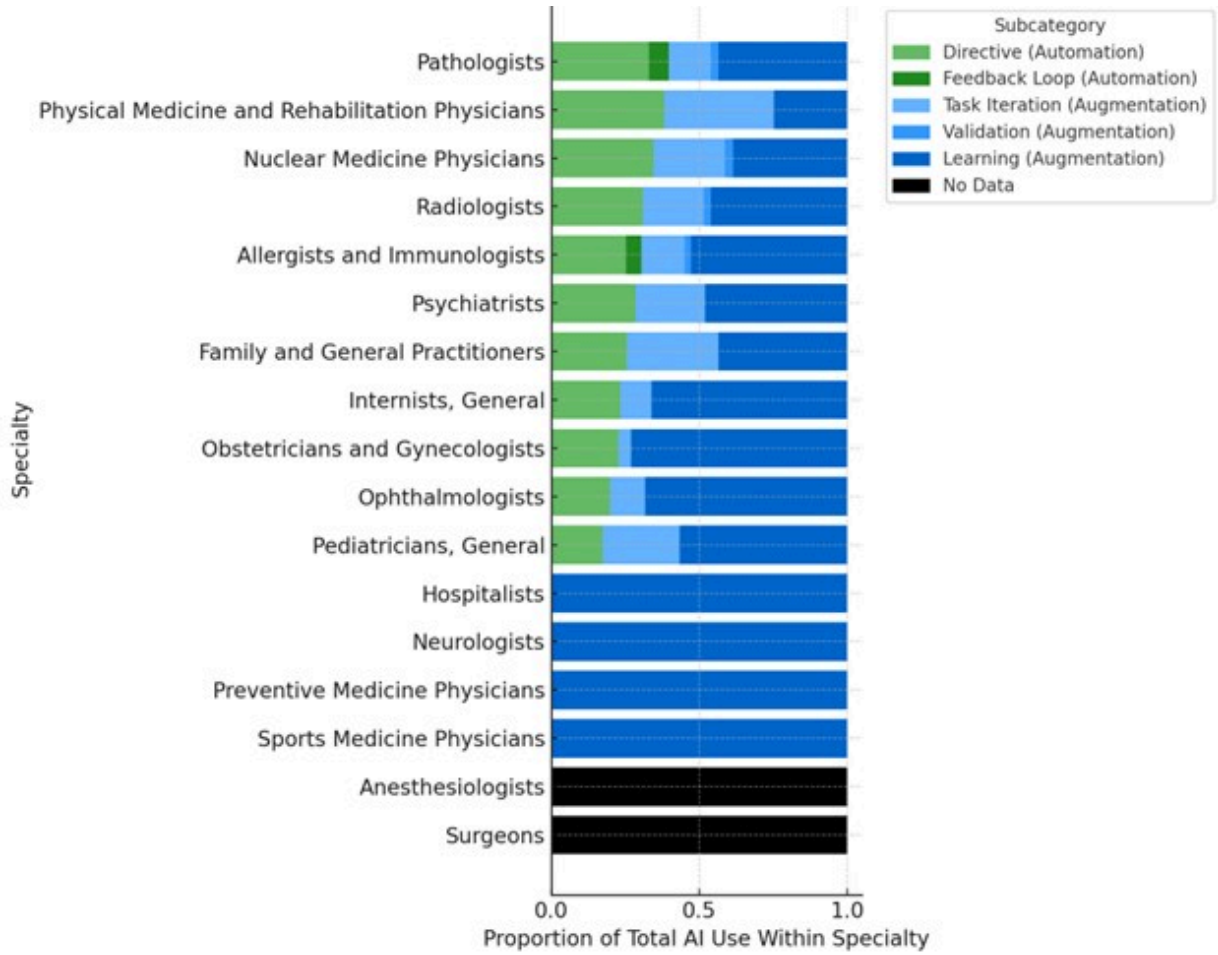

Figure 1

Type of user interaction by medical specialty

*Note: No user interaction data reported for anesthesiologists and surgeons. Bars represent within-specialty proportions that sum to 1; therefore, stacked segment sizes reflect the relative distribution of user interaction patterns within each specialty rather than total volume across all specialties.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- economicindexappendix200226.docx