Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Perceived impact of generative artificial intelligence on clinical learning and decision-making among internal medicine residents in Ethiopia: a mixed- methods study

Hayatu Awel Abdela

hayatu.awol@wku.edu.et

Wolkite University

Tamirat Godebo Woyimo

Jimma University

Kenzudine Assfa Mossa

Wolkite University

Research Article

Keywords: generative artificial intelligence, large language models, medical education, clinical reasoning, internal medicine residency, Ethiopia, mixed methods

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11541959/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Generative artificial intelligence (GenAI) is increasingly used in medical training, but evidence from African residency programs is scarce. We assessed how internal medicine residents in Ethiopia use GenAI and how they perceive its impact on clinical learning and clinical decision-making.

Methods

We conducted an explanatory sequential mixed-methods study consisting of a national cross-sectional online survey of internal medicine residents (June to August 2026), followed by two focus group discussions (FGDs) with 10 residents at two universities (September 2026) to explain the survey findings. Perceived impact was measured with five-point Likert items and two composite scores (learning, five items; decision-making, eight items). Associations with frequency of use were tested with chi-square or Fisher's exact tests. FGD data were organized deductively into categories based on the survey topics.

Results

Of 252 respondents from 18 institutions, 98.8% had used GenAI and 75.4% used it daily; ChatGPT was used by 91.6% of users. The mean composite score for perceived impact on clinical learning was 3.83 out of 5 (70.7% high) and for clinical decision-making 3.86 out of 5 (67.7% high). Agreement was highest for improved differential diagnosis (90.8%), learning efficiency (81.9%) and support for self-directed learning (79.5%), and lowest for ward-round performance (64.3%) and confidence in complex cases (71.4%). Daily use was more frequent among residents with more than three hours of daily internet use (82.3% versus 66.7% to 67.4%; $p = 0.024$) and was not associated with age, sex, marital status, year of residency, device access or formal training. In the FGDs, residents described GenAI as a fast, supportive tool for learning and clinical work, as an aid to reasoning and not a replacement for physician judgment, and as a possible threat to independent learning.

Conclusions

Residents perceive GenAI as helpful for self-directed learning and diagnostic reasoning while keeping final responsibility for clinical decisions. These findings reflect perceptions and not measured outcomes, so objective evaluation of its effects on learning and clinical performance is needed.

Background

Generative artificial intelligence (GenAI), particularly large language models (LLMs), can explain concepts, summarize large bodies of text and suggest differential diagnoses on demand, and interest in its use in medical education has grown quickly [1–3]. Reviews describe its promise for personalized learning and rapid access to information [3]. Residency is a setting where this promise is easy to understand. Residents must integrate large amounts of rapidly changing knowledge with diagnostic reasoning, treatment planning and patient care, often under time pressure.

Whether GenAI actually improves clinical reasoning is less certain. In a randomized trial, access to an LLM did not improve physicians' diagnostic reasoning compared with conventional resources [4], whereas in another, GPT-4 assistance improved performance on management-reasoning tasks [5]. In a large evaluation using real patient cases, LLMs performed worse than physicians and did not reliably follow guidelines [6]. Learners' own perceptions of benefit therefore need to be interpreted alongside this evidence. Surveys of residents are still few; in one from North Carolina, 26% of internal medicine residents reported current use of LLMs [7].

In low- and middle-income countries, the evidence comes mostly from students. Use of ChatGPT and other AI tools is widespread among medical students in Uganda [8], and Egyptian students reported strong interest alongside concerns about reliability [9]. In Ethiopia, health professionals recognize the potential of AI but identify barriers related to infrastructure, knowledge, trust and ethics [10]. To our knowledge, no study has described GenAI use and its perceived impact on learning and decision-making among internal medicine residents in Africa.

We therefore conducted a mixed-methods study to assess how internal medicine residents in Ethiopia use GenAI and how they perceive its impact on clinical learning and clinical decision-making. Our specific objectives were to describe the prevalence, frequency, tools and situations of use; to assess perceived impact on clinical learning; to assess perceived impact on clinical decision-making; to identify factors associated with frequency of use; and to explore, through focus group discussions (FGDs), how residents use GenAI to learn and reason clinically and how they distinguish support from decision-making.

Methods

Study design

We used an explanatory sequential mixed-methods design [11]. An online cross-sectional survey was conducted first, and focus group discussions (FGDs) followed to explain and extend its findings. The survey measured GenAI use and its perceived impact on clinical learning and decision-making. The FGDs explored how residents use GenAI and why they hold their views, and their guide was prepared after the preliminary survey results had been reviewed. The two components were analyzed separately and brought together at the interpretation stage. Findings on trust in GenAI, verification practices, barriers, and institutional readiness from the same survey and focus group discussions are presented separately

in a companion manuscript. The survey is reported according to the STROBE statement [12] and the FGDs according to the COREQ checklist [13].

Setting and participants

We included residents in years 1 to 3 (R1 to R3) of internal medicine residency programs in Ethiopia, estimated at 450 to 500 in 2026. The survey ran from June to August 2026 and the FGDs were held in September 2026.

Survey recruitment. The questionnaire link was posted in the Telegram group of each internal medicine department. Every resident belongs to this group, which is the department's usual channel for announcements. Completing the form was voluntary, and because the initial response was slow, departments and chief residents sent reminders at intervals. Residents who consented and completed the questionnaire were included. Because the link was shared in group chats, we could not determine how many residents saw it, and the reasons for non-participation were not assessed. In total, 252 residents from 18 institutions (15 universities, one medical college and two hospitals) responded, about 50% to 56% of the estimated population.

FGD recruitment. Two institutions in different settings were purposively selected to represent different generations of residency programs: Addis Ababa University (AAU), a first-generation program in the capital city, and Wolkite University (WKU), a third-generation program in a regional town. The invitation was posted in the same departmental Telegram group, so participants already knew about the survey. Residents who voluntarily attended the department hall at the announced time took part. The number of residents who saw the invitation but did not attend was not recorded; no participants left either session before its conclusion. Each FGD had five participants, from all three residency years. One participant was a woman (AAU); there were no women at WKU.

Sample size

The minimum survey sample was estimated with Cochran's formula for a proportion [14], using a 95% confidence level, an expected proportion of 0.5 and a 5% margin of error, which gives 384. With the finite-population correction for about 500 residents, the minimum was approximately 218. We aimed for at least 250 completed questionnaires to allow for non-response and subgroup comparisons, and obtained 252. Two FGDs were planned because their purpose was to explain the survey findings, not to reach thematic saturation.

Data collection

Survey. The questionnaire was developed by the research team from the literature on GenAI in medical education and administered through Google Forms; it took about five minutes to complete. It was pre-

tested with 15 residents who were not in the final sample and revised accordingly. It was not formally validated. For this paper we used these sections: socio-demographic characteristics (age, sex, marital status, residency year, institution); digital access (computer, laptop or tablet; smartphone use at the point of care; electricity interruptions; daily internet use); GenAI use (frequency, tools, and educational and clinical situations of use); prior formal GenAI training (yes or no); and perceived impact, measured with twelve five-point Likert items (strongly disagree to strongly agree). Internal consistency was good, with a Cronbach's alpha of 0.867 for the five-item learning scale and 0.839 for the eight-item decision-making scale.

FGDs. The FGDs were led by two authors, both men. KAM (BSc, MPH; public health, with experience in qualitative research) moderated the discussions. HAA (MD; consultant internist) co-facilitated. Neither author had any link with the AAU residents. At WKU, KAM had no link with the residents, whereas HAA mentors them. Each session was held in the department hall, with only the two authors and the participating residents present. The facilitators introduced themselves and explained the aim of the study and of the discussion, which was to explain the survey results and to explore any ideas the survey had not covered. Participants were told that there were no right or wrong answers and that their views would be reported anonymously.

The semi-structured guide was drafted by the authors after reviewing the preliminary survey results and was not pre-tested. It covered views on AI in internal medicine, frequency of use and tools, effects on learning and clinical decision-making, perceptions of institutional support and readiness, training needs, barriers, facilitators and recommendations for integration. One session was held at each site and no repeat discussions were held. Participants could speak Amharic, English or both. With permission, the sessions were audio-recorded, and both authors took field notes. The recordings were transcribed manually and verbatim into English, keeping the intended meaning where the wording was unclear. The AAU discussion lasted 56 minutes and the WKU discussion 68 minutes. Transcripts were not returned to participants.

Variables

GenAI use refers to use of generative AI tools, including LLM chatbots such as ChatGPT and AI-enabled clinical reference tools such as UpToDate AI. A GenAI user had used such a tool at least once. Frequency of use was dichotomized as daily or less than daily among users. Perceived impact on clinical learning was a composite of five items (learning efficiency, self-directed learning, examination preparedness, ward-round performance and overall positive influence on clinical learning), calculated as the mean item score (range 1 to 5; mean sum 5 to 25). Perceived impact on clinical decision-making was a composite of eight items (diagnostic accuracy, clinical reasoning, decision-making time, differential diagnosis, treatment planning, confidence in complex cases, interpretation of investigations and overall positive influence), calculated in the same way (mean sum 8 to 40). Both composites were also grouped into equal thirds of the five-point range: low (1 to 2.33), moderate (2.34 to 3.67) and high (3.68 to 5).

Explanatory variables for frequency of use were age, sex, marital status, year of residency, device access, daily internet use and formal training.

Data analysis

Survey. Data were analyzed with IBM SPSS Statistics, version 24. We report frequencies, percentages with 95% confidence intervals (Wilson method) and means; agreement means "agree" or "strongly agree". Missing data were few (1 to 4 per item) and were not imputed, and valid sample sizes are shown. No weighting was applied because all residents were invited. Associations between frequency of use (daily versus less than daily, among users) and participant characteristics, digital access and training were examined with the chi-square test, or Fisher's exact test when expected counts were small, and summarized with unadjusted odds ratios and 95% confidence intervals. Significance was set at a two-sided $p < 0.05$. These analyses were exploratory, were not adjusted for confounding or multiple comparisons, and are not interpreted as causal. To limit selection bias we invited all residents and sent reminders, and we pre-tested the questionnaire to reduce measurement error; the remaining biases are discussed in the limitations.

FGDs. The transcripts from all 10 FGD participants were coded manually by KAM, without qualitative data analysis software, using a deductive approach consistent with directed content analysis [15]. Categories were predefined based on the survey topics, and participants' statements were assigned to these categories. Statements that did not fit the predefined categories, including those concerning conflicts with local protocols and patient privacy, were considered separately. To check coding consistency, the transcripts were cross-checked against both authors' field notes and the audio recordings. Participants did not review the findings. Quotations are identified by study site, participant code, and residency year.

Integration. Survey and FGD findings were compared for each objective and are shown in a joint display (Table 7).

Ethical considerations

Ethical approval

was obtained from the Institutional Review Board of Wolkite University (reference number: 984/05/2026; approval date: 14 May 2026). All participants gave informed consent before taking part, electronically for the survey and before each FGD, including consent for audio recording. Participation was voluntary. The study was conducted in accordance with the Declaration of Helsinki. Data were handled securely and de-identified, and FGD participants are identified only by codes.

Results

Participants and patterns of GenAI use

In total, 252 residents from 18 institutions completed the survey, about 50% to 56% of the estimated population. The largest contributions came from Addis Ababa University (13.5%), Jimma University (11.1%) and Wolkite University (9.9%). Most respondents were aged 31 to 40 years (67.1%), male (88.1%) and in their third year of residency (46.8%) (Table 1). Ten residents took part in the FGDs, five at each site, from all three residency years; one was a woman.

Table 1
Characteristics of survey respondents (N = 252)

Variable	Category	n (%)
Age (years)	24–30	76 (30.2)
	31–40	169 (67.1)
	> 40	7 (2.8)
Sex	Male	222 (88.1)
	Female	30 (11.9)
Marital status	Single	126 (50.0)
	Married	123 (48.8)
	Divorced	3 (1.2)
Residency year	R1	66 (26.2)
	R2	68 (27.0)
	R3	118 (46.8)

GenAI use was almost universal. Overall, 249 residents (98.8%) had used GenAI, 231 (91.7%) used it at least weekly and 190 (75.4%) used it daily (Table 2). Digital access was high: 96.4% had a computer, laptop or tablet, 98.4% used a smartphone at the point of care and 59.1% used the internet for more than three hours daily. ChatGPT dominated, being used by 91.6% of users, followed by UpToDate AI (36.9%), Grok (14.5%), NotebookLM (11.2%) and Claude (8.4%). The most common situations of use were differential diagnosis (70.3%), studying (67.9%), research (60.2%), examination preparation (60.2%) and drug information (57.8%). GenAI was also used in clinical environments, including outpatient clinics (45.8%), emergency care (43.8%), laboratory interpretation (40.6%), radiology (37.8%), clinical rounds (34.5%) and ECG interpretation (30.5%) (Fig. 1).

Table 2
Frequency of GenAI use and tools used

Characteristic	n/N	% (95% CI)
Frequency of use (N = 252)		
Never	3/252	1.2 (0.4–3.4)
Rarely	15/252	6.0 (3.6–9.6)
Monthly	3/252	1.2 (0.4–3.4)
Weekly	41/252	16.3 (12.2–21.3)
Daily	190/252	75.4 (69.7–80.3)
Tools used (among 249 users; multiple responses)		
ChatGPT	228/249	91.6 (87.5–94.4)
UpToDate AI	92/249	36.9 (31.2–43.1)
Grok	36/249	14.5 (10.6–19.4)
NotebookLM	28/249	11.2 (7.9–15.8)
Claude	21/249	8.4 (5.6–12.5)
Gemini	14/249	5.6 (3.4–9.2)
DeepSeek	8/249	3.2 (1.6–6.2)
Perplexity	2/249	0.8 (0.2–2.9)

Perceived impact on clinical learning

Residents perceived a favorable impact of GenAI on clinical learning (Table 3). The composite mean score was 3.83 out of 5 (mean sum 19.14 out of 25). Learning efficiency (82%), overall positive influence on clinical learning (82.3%) and support for self-directed learning (79.5%) had the highest agreement, followed by examination preparedness (74.3%). Ward-round performance had the lowest agreement (64.3%; mean 3.64), although a clear majority still agreed. Based on the composite score, 176 residents (70.7%) had a high perceived impact, 69 (27.7%) a moderate impact and 4 (1.6%) a low impact.

Table 3
Perceived impact of GenAI on clinical learning

Item	Valid n	Mean (of 5)	Agree n	% (95% CI)
Improves learning efficiency	249	3.95	204	82 (76.7–86.2)
Supports self-directed learning	249	3.90	198	79.5 (74.1–84.1)
Improves examination preparedness	249	3.78	185	74.3 (68.5–79.3)
Improves ward-round performance	249	3.64	160	64.3 (58.1–70.0)
Positive influence on clinical learning (overall)	249	3.88	205	82.3 (77.1–86.6)
Composite score (mean of five items)	249	3.83	–	–

Agree = agree or strongly agree. CI, confidence interval (Wilson).

Perceived impact on clinical decision-making

Perceived impact on clinical decision-making was also favorable (Table 4), with a composite mean score of 3.86 out of 5 (mean sum 30.90 out of 40). Improved differential diagnosis was the most strongly endorsed item (90.8%; mean 4.08), followed by clinical reasoning (83.1%), interpretation of investigations (82.7%), treatment planning (79.1%), overall decision-making (77.4%), reduced decision-making time (76.3%) and diagnostic accuracy (75.0%). Confidence in managing complex cases had the lowest mean (3.69), although 71.4% agreed. Based on the composite score, 168 residents (67.7%) had a high perceived impact, 76 (30.6%) a moderate impact and 4 (1.6%) a low impact (n = 248).

Table 4
Perceived impact of GenAI on clinical decision-making

Item	Valid n	Mean (of 5)	Agree n	% (95% CI)
Improves differential diagnosis	249	4.08	226	90.8 (86.5–93.8)
Enhances clinical reasoning	249	3.95	207	83.1 (78.0–87.3)
Improves interpretation of investigations	249	3.93	206	82.7 (77.5–86.9)
Improves treatment planning	249	3.86	197	79.1 (73.6–83.7)
Improves diagnostic accuracy	248	3.83	186	75.0 (69.3–80.0)
Positive influence on clinical decision-making (overall)	248	3.83	192	77.4 (71.8–82.2)
Reduces decision-making time	249	3.78	190	76.3 (70.6–81.2)
Improves confidence in complex cases	248	3.69	177	71.4 (65.5–76.6)
Composite score (mean of eight items)	248	3.86	–	–

Agree = agree or strongly agree. CI, confidence interval (Wilson).

Factors associated with frequency of GenAI use

Among the 249 GenAI users, 190 (76.3%) used GenAI daily and 59 (23.7%) less often **(**Table 5**).** Frequency of use was not associated with age, sex, marital status, year of residency, device access or formal training. Daily use was more common among residents who used the internet for more than three hours a day (82.3%) than among those using it for less than one hour (66.7%) or one to three hours (67.4%) (p = 0.024). Compared with residents who used the internet for three hours or less, those who used it for more than three hours had higher odds of daily use.

Table 5
Factors associated with daily versus less-than-daily GenAI use among GenAI users (N = 249)

Variable	Category	Daily use n (%)	Less than daily n (%)	p-value
Age (years)	24–30	56 (74.7)	19 (25.3)	0.867
	31–40	129 (77.2)	38 (22.8)	
	> 40	5 (71.4)	2 (28.6)	
Sex	Male	171 (77.7)	49 (22.3)	0.146
	Female	19 (65.5)	10 (34.5)	
Marital status	Single	100 (80.0)	25 (20.0)	0.376
	Married	88 (72.7)	33 (27.3)	
	Divorced	2 (66.7)	1 (33.3)	
Residency year	R1	50 (76.9)	15 (23.1)	0.988
	R2	50 (75.8)	16 (24.2)	
	R3	90 (76.3)	28 (23.7)	
Computer, laptop or tablet access	No	6 (75.0)	2 (25.0)	1.000†
	Yes	184 (76.3)	57 (23.7)	
Daily internet use	< 1 hour	10 (66.7)	5 (33.3)	0.024
	1–3 hours	58 (67.4)	28 (32.6)	
	> 3 hours	121 (82.3)	26 (17.7)	
Formal GenAI training	No	182 (76.5)	56 (23.5)	0.705†
	Yes	7 (70.0)	3 (30.0)	

†Fisher's exact test; other p-values from the chi-square test.

Qualitative findings

Three categories from the FGDs relate to how residents use GenAI to learn and reason clinically (Table 6). Other categories, on trust, verification and institutional readiness, are reported separately.

GenAI as a rapid and supportive tool for learning and clinical work

Participants consistently described GenAI as useful for fast information retrieval, summarizing lengthy materials, explaining difficult concepts, preparing seminars and examinations, and reducing workload, especially when time is short. They also described using it for differential diagnosis, laboratory

interpretation, drug information, unfamiliar conditions and quick revision before clinical rounds, and in the emergency department to identify unfamiliar poisoning cases. As one resident said, “Even just before a clinical round, you can ask AI to get a fast and on point answer” (FGD-AAU-P2). Another said, “AI will save our time, we will get immediate response, it simplifies complex ideas” (FGD-WKU-P5).

GenAI as an aid to clinical reasoning, not a replacement for physician judgment

Participants valued GenAI for broadening differential diagnoses, identifying unfamiliar diseases, reminding them of diagnostic criteria and supporting reasoning in complex cases. They distinguished clearly between support and decision-making. The treating physician, they said, knows the full clinical context and must keep responsibility for diagnosis and management. Two participants said they never used GenAI to make clinical decisions. One described a case in which the output prompted further reading: “I have encounter unconsidered/unusual differential diagnosis from AI, and help me to read about that new/rare disease conditions” (FGD-AAU-P3). Another summarized the position: “AI can assist us, not replaces the physician clinical judgements” (FGD-WKU-P2).

Risk of overdependence and reduced independent learning

Participants worried that rapid answers could reduce independent reading, searching, retention and reasoning, and could encourage dependence for routine academic tasks. Several stressed that textbooks and conventional learning remain important and that GenAI should supplement them. One said, “The drawback is, it decreases our efforts to search more, since we are students we will learn while searching, but in using AI it reduces this” (FGD-WKU-P1). Another said, “There is nothing left in our mind, as if I’m using AI as my brain” (FGD-WKU-P5). Participants therefore saw the manner, purpose and extent of use as decisive in whether GenAI supports or undermines learning.

Table 6
Qualitative categories relevant to learning and clinical decision-making

Category	Key topics	Illustrative quotation
GenAI as a rapid, supportive tool for learning and clinical work	Rapid retrieval; explaining difficult concepts; summarizing; seminar and examination preparation; differential diagnosis; workload reduction	“Even just before a clinical round, you can ask AI to get a fast and on point answer.” (FGD-AAU-P2)
Support rather than replacement in clinical decision-making	Additional differentials; diagnostic support; reluctance to delegate final decisions; physician responsibility	“AI can assist us, not replaces the physician clinical judgements.” (FGD-WKU-P2)
Risk of overdependence and reduced independent learning	Less searching and reading; dependency; impaired retention and independent reasoning	“There is nothing left in our mind, as if I’m using AI as my brain.” (FGD-WKU-P5)

Discussion

Principal findings

In this national mixed-methods study, almost all internal medicine residents used GenAI and three-quarters used it daily. Table 7 summarizes how the survey and FGD findings fit together. Most perceived a high impact on both clinical learning and clinical decision-making, with the strongest agreement for differential diagnosis, learning efficiency and self-directed learning and the weakest for ward-round performance and confidence in complex cases. The FGDs explain these patterns: residents use GenAI for speed, treat it as a reasoning aid and not a decision-maker, and worry about what rapid answers do to independent learning.

Table 7
Joint display of survey and focus group findings

Survey finding	FGD explanation	Integrated interpretation
Almost universal use (98.8%); daily use 75.4%	Time pressure, rapid access and workload reduction	Practical utility drives adoption
High perceived learning impact (composite 3.83 of 5); strongest for learning efficiency (81.9%) and self-directed learning (79.5%)	Concept explanation, summarizing, seminar and examination preparation, quick revision	GenAI mainly supports individual study
Ward-round performance rated lowest (64.3%)	GenAI lacks the full patient context and local protocols	Less useful at the bedside than for study
High perceived decision-making impact (composite 3.86 of 5); differential diagnosis 90.8%; confidence in complex cases lowest (3.69)	Unusual differentials and prompts for further reading; residents keep final responsibility	GenAI is used as an adjunct to reasoning, not a replacement for judgment
Favorable learning scores	Less searching and reading; worry about dependence and retention	Perceived benefit depends on how GenAI is used

Use of GenAI

Use was far higher than in the closest comparable survey of internal medicine residents, in which 26% reported current use [7]. Part of the difference is probably timing, since that survey was published in 2024 and GenAI tools have spread quickly since, and part may reflect differences in setting, sampling and definitions. Widespread use has also been reported among medical students in Uganda [8]. Several features of the Ethiopian setting may help explain the uptake, although we did not test them: device access was almost universal, the main tools are free, and the FGDs point to time pressure and workload as practical drivers. Daily use was more common among residents with more than three hours of daily internet use, which fits an access explanation. Because respondents chose to take part, and more digitally engaged residents may have been more likely to respond, true prevalence among all residents may be somewhat lower. Use was also concentrated on general-purpose tools, mainly ChatGPT, which were not designed for clinical use, with UpToDate AI a distant second.

Perceived impact on clinical learning

Residents rated GenAI most highly for learning efficiency and self-directed learning and least for ward-round performance. This is consistent with reviews that describe personalized learning and rapid access to information as the main educational promise of LLMs [1–3]. The gradient is plausible. GenAI works on text and explanation, so it suits tasks that residents carry out alone, such as clarifying a concept, condensing reading or preparing for an examination. Ward rounds depend on history taking, examination, communication and local protocols, which a text-based tool cannot do directly. The FGDs support this reading, with residents describing concept explanation, summarizing, seminar preparation and quick revision before rounds, and noting that GenAI lacks the full patient context.

The favorable scores are only part of the picture. The FGDs showed that residents recognize a cost: less searching, less reading and weaker retention. Some described feeling that they use AI “as my brain”. Residents therefore see the educational value of GenAI as conditional on how it is used. These are perceptions, and none of the studies we cite, including ours, shows that GenAI improves knowledge retention or examination performance.

Perceived impact on clinical decision-making

Differential diagnosis was the most strongly endorsed benefit (90.8%), followed by clinical reasoning and interpretation of investigations. Perceived benefit should be set against experimental evidence, which is mixed: LLM access did not improve physicians' diagnostic reasoning in one trial [4] but improved management reasoning in another [5], and in real patient cases LLMs performed worse than physicians and did not reliably follow guidelines [6]. The high rating for differential diagnosis likely reflects a task at which these tools are strong, which is generating options, and describes an experience of being helped. Our study did not measure whether diagnostic accuracy improved.

The FGDs add a distinction that the questionnaire cannot capture, between GenAI that supports reasoning and GenAI that makes the decision. Residents used it to widen the differential and to prompt further reading, but kept responsibility for the final decision, and two said they never used it for clinical decisions. This helps to reconcile the high survey scores with the cautious tone of the discussions: agreement that GenAI improves decision-making can reflect brainstorming and reminding, which residents distinguished sharply from delegating a decision. The lower score for confidence in complex cases fits the same picture, since complex cases are those where missing patient context matters most. The findings suggest that residents largely use GenAI as a reasoning aid, which is consistent with evidence that current models are not suitable for autonomous decisions [6], but the boundary between aid and substitute is held by individual judgment and not by institutional guidance.

Implications

Residency programs can use GenAI's perceived strength in self-directed learning deliberately, for example in concept review and examination preparation, while teaching residents to treat output as a

starting point for reading and not as an answer. Training should emphasize the line between support and decision, and the risk of dependence. Educators and researchers should test the educational effects objectively, with measures such as knowledge retention, examination performance and diagnostic accuracy in the same residents. Trust, verification, and institutional readiness for safe GenAI integration are examined in a companion manuscript currently submitted to BMC Medical Education.

Strengths and limitations

Strengths include national coverage of 18 institutions, the combination of survey and FGD data, and the use of composite scores alongside individual items. Several limitations apply. The cross-sectional design does not allow causal inference, and all impact measures are self-reported perceptions and not measured learning or clinical performance. Most benefit items were worded positively and rated on agreement scales, which may encourage agreement, and the questionnaire was not formally validated. The composite categories (low, moderate, high) are equal divisions of the scale with no established thresholds. Although all residents belong to their department's Telegram group, participation was voluntary and depended on reminders, so residents who were more engaged with digital tools may have been more likely to respond, which may overestimate use; the response rate was approximately 50% to 56% of the estimated population. The findings describe internal medicine residents in Ethiopia who responded to an online survey and may not generalize to other specialties, countries or non-respondents. At Wolkite University, one of the facilitators (HAA) mentors the residents, which may have influenced responses despite assurances of an open discussion. Most respondents were men (88.1%), which may reflect the sex distribution of internal medicine residents in Ethiopia, but the small number of women (n = 30) precluded sex-specific analysis. The qualitative component comprised two FGDs of volunteer residents at two universities, with only one woman among the ten participants and a guide that was not pre-tested, so the findings are explanatory and not representative and saturation cannot be assumed. The associations with frequency of use were bivariate, exploratory and not adjusted for multiple testing or confounding. GenAI tools are evolving rapidly, and these findings may date quickly.

Conclusions

Internal medicine residents in Ethiopia use GenAI almost universally and perceive it as helpful for self-directed learning and diagnostic reasoning, while keeping final responsibility for clinical decisions and recognizing a risk to independent learning. Because these are perceptions, objective evaluation of the effects of GenAI on learning and clinical performance should be the next step, alongside structured guidance on its appropriate use in residency education.

Declarations

Ethics approval and consent to participate:

Approved by the Institutional Review Board of Wolkite University (reference number: 984/05/2026; approval date: 14 May 2026). All participants provided informed consent.

Consent for publication:

Not applicable. No identifiable individual data are reported.

Competing interests:

The authors declare no competing interests.

Funding:

This study was supported by a competitive research grant of ETB 100,000 from the Ethiopian Society of Internal Medicine (ESIM), awarded following a call for research proposals on artificial intelligence. The funder had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Author Contribution

Authors' contributions: HAA contributed to conceptualization, funding acquisition, investigation, supervision, and writing of the original draft. TGW contributed to methodology, investigation, and review and editing of the manuscript. KAM contributed to qualitative data collection and analysis, and review and editing of the manuscript. All authors read and approved the final manuscript.

Acknowledgement

Acknowledgements: We thank the residents who participated, the program coordinators who facilitated data collection, and the Ethiopian Society of Internal Medicine.

Data Availability

Availability of data and materials: The de-identified survey dataset and questionnaire are available from the corresponding author upon reasonable request. Qualitative transcripts are not publicly available to protect participant confidentiality.

References

1. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME Guide 84. Med Teach. 2024;46(4):446–70. https://doi.org/10.1080/0142159X.2024.2314198.
2. Tozsin A, Ucmak H, Soyturk S, Aydin A, Gozen AS, Al Fahim M, et al. The role of artificial intelligence in medical education: a systematic review. Surg Innov. 2024;31(4):415–23. https://doi.org/10.1177/15533506241248239.
3. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024;58(11):1276–85. https://doi.org/10.1111/medu.15402.
4. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969.
5. Goh E, Gallo RJ, Strong E, Weng Y, Kerman H, Freed JA, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. Nat Med. 2025;31:1233–8. https://doi.org/10.1038/s41591-024-03456-y.
6. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613–22. https://doi.org/10.1038/s41591-024-03097-1.
7. Fried AJ, Dorn SD, Leland WJ, Mullen E, Williams DM, Zaas AK, et al. Large language models in internal medicine residency: current use and attitudes among internal medicine residents. Discov Artif Intell. 2024;4:70. https://doi.org/10.1007/s44163-024-00173-w.
8. Ajalo E, Mukunya D, Nantale R, Kayemba F, Pangholi K, Babuya J, et al. Widespread use of ChatGPT and other artificial intelligence tools among medical students in Uganda: a cross-sectional study. PLoS ONE. 2025;20(1):e0313776. https://doi.org/10.1371/journal.pone.0313776.
9. Abdelhafiz AS, Farghly MI, Sultan EA, Abouelmagd ME, Ashmawy Y, Elsebaie EH. Medical students and ChatGPT: analyzing attitudes, practices, and academic perceptions. BMC Med Educ. 2025;25(1):187. https://doi.org/10.1186/s12909-025-06731-9.
10. Kasaye MD, Getahun AG, Kalayou MH. Exploring artificial intelligence for healthcare from the health professionals' perspective: the case of limited resource settings. Digit Health. 2025;11:20552076251330552. https://doi.org/10.1177/20552076251330552.
11. Creswell JW, Plano Clark VL. Designing and conducting mixed methods research. 3rd ed. Thousand Oaks (CA): SAGE; 2018.
12. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453–7. https://doi.org/10.1016/S0140-6736(07)61602-X.
13. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care. 2007;19(6):349–57.

14. Cochran WG. Sampling techniques. 3rd ed. New York: Wiley; 1977.
15. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res. 2005;15(9):1277–88. https://doi.org/10.1177/1049732305276687.

Figures

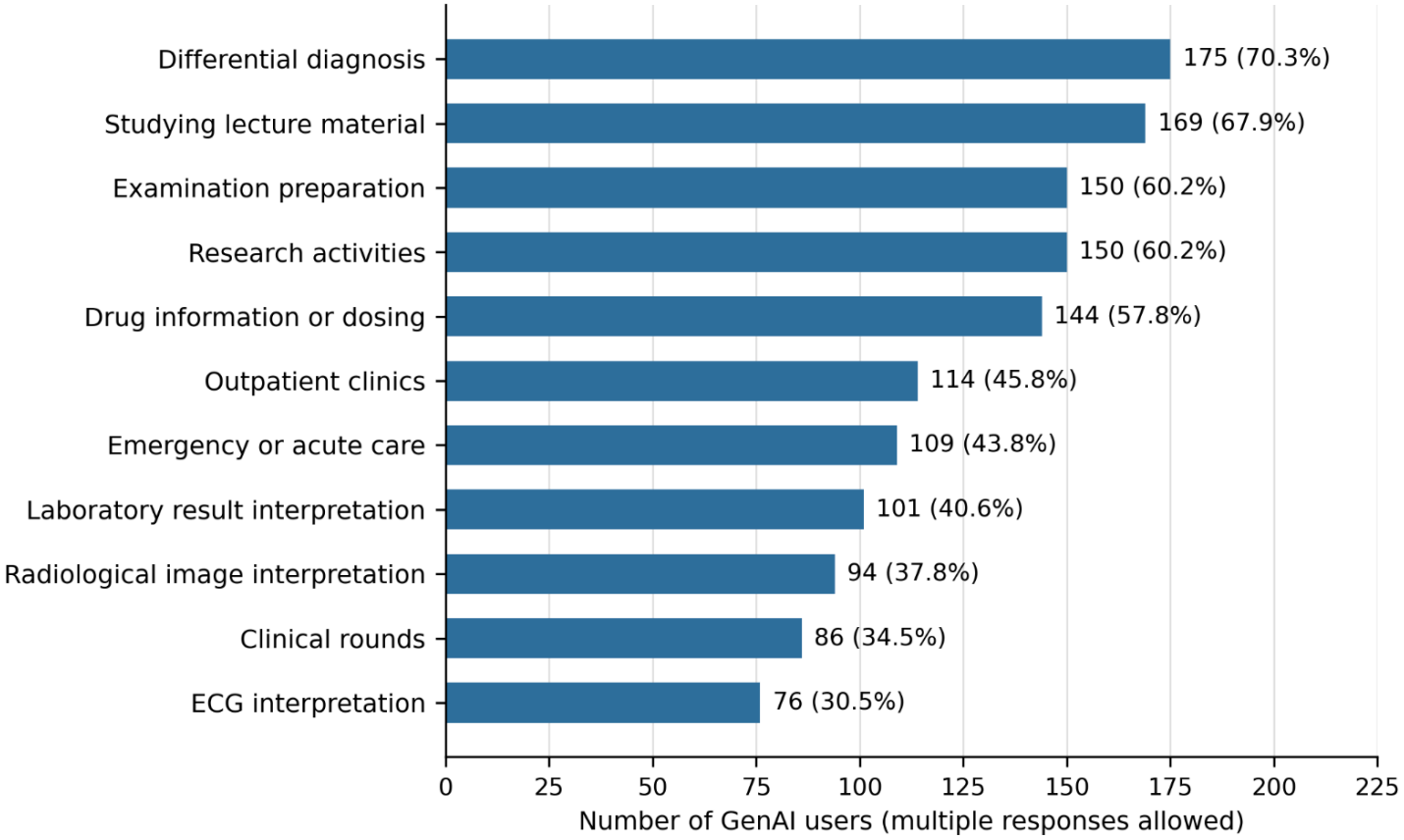


Figure 1

Situations in which GenAI was used among GenAI users (n = 249; multiple responses allowed)

Supplementary Files

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

- Supplimentarydoc1FGDsGuide.docx
- Supplimentarydoc3COREQChecklistManuscript1.docx
- Supplimentarydoc2STROBEChecklistManuscript1.docx