Clinical Imaging 125 (2025) 110525

Contents lists available at ScienceDirect

Clinical Imaging

journal homepage: www.elsevier.com/locate/clinimag

Breast Imaging

Classifying the clinical significance of common breast pain symptoms using a large language model, ChatGPT (GPT-4)

Hana Haver[a], Manisha Bahl[b], Maggie Chung[c,*]

[a] Department of Radiology, Massachusetts General Brigham, Boston, MA, United States of America
[b] Department of Radiology, Massachusetts General Hospital, Boston, MA, United States of America
[c] Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, United States of America

The American College of Radiology (ACR) Appropriateness Criteria[1] defines clinically insignificant breast pain as nonfocal, diffuse, or cyclical pain. Clinically insignificant pain does not require diagnostic imaging evaluation because it is not associated with malignancy.[1] Diagnostic mammograms and/or ultrasound examinations performed for clinically insignificant pain often lack imaging correlates.[2,3] In the context of high workflow demands and prolonged patient wait times, strategies to minimize unnecessary diagnostic imaging for clinically insignificant pain include having nursing staff contact patients to confirm pain symptoms and reschedule unwarranted diagnostic imaging appointments as age-appropriate screening appointments. Large language models (LLMs) have been evaluated for various radiology applications, such as providing patient education regarding breast cancer prevention and screening[4] and offering clinical decision support for clinicians managing patients with breast pain.[5] The purpose of this study was to evaluate the performance of LLM, ChatGPT GPT-4 (March 2023 release, OpenAI) in automating the classification of patients with breast pain as either clinically significant and therefore requiring diagnostic imaging evaluation or clinically insignificant and not requiring imaging evaluation.

This retrospective study received an institutional review board exemption. For this study, one breast imaging fellow (H.H.) and two fellowship-trained breast imaging radiologists (M.C., 1 year of post-training experience, and M.B., 8 years of post-training experience) developed 150 patient-centered breast pain clinical vignettes encompassing clinical variants described in the ACR Appropriateness Criteria for Breast Pain[1] and frequently encountered patient symptoms based on authors' clinical experience (Supplemental Table). Thirty of the 150 breast pain symptoms (15 clinically insignificant and 15 clinically significant pain) also included non-pain clinically significant symptoms (e.g., palpable lump, pathologic nipple discharge). The ground truth for the clinical significance of each symptomatic variant was determined based on consensus by the authors.

A zero-shot prompt, which asks GPT-4 to generate responses without prior examples or specific training, was developed for the purpose of this study, to characterize breast symptoms as clinically insignificant or clinically significant in accordance with the ACR Appropriateness Criteria for Breast Pain: "Use the ACR appropriateness criteria for breast pain. Respond with only is this 'clinically significant breast symptom' or 'not clinically significant symptom.'" Each breast pain symptom was submitted with the same prompt in three independent tests, performed on three different computers, in June 2024. All responses from ChatGPT GPT-4 were recorded as "clinically significant" or "not clinically significant," and the final ChatGPT GPT-4-assigned designated response of clinical significance was determined by the mode of the three tests. Assigned clinical significance from ChatGPT GPT-4 was compared to the ground truth established by breast radiologist consensus.

Based on radiologist consensus, of the 150 clinical vignettes, 64 (64/150; 42.7 %) were considered clinically significant and 86 (86/150; 57.3 %) were assessed as clinically insignificant (Table 1). Of the 64 clinically significant cases, 34 vignettes described only significant pain alone and 30 were pain symptoms that also included non-pain clinically significant symptoms (e.g., palpable lump, pathologic nipple discharge).

ChatGPT GPT-4 correctly classified 112 of 150 (74.7 %) of clinical vignettes with breast pain symptoms (Table 1). Of the 64 cases with either clinically significant pain or an additional concerning symptom (e.g. palpable concern, nipple discharge), ChatGPT GPT-4 correctly identified 57 cases for a sensitivity of 89.1 % and a false negative rate of 10.9 %. All 30 pain symptoms that also included non-pain clinically significant symptoms (e.g., palpable lump, pathologic nipple discharge) were correctly classified as clinically significant by ChatGPT GPT-4. Of the 86 clinical vignettes with clinically insignificant symptoms, ChatGPT GPT-4 correctly identified 55 cases for a specificity of 64.0 %. Among those symptoms that were misclassified by ChatGPT GPT-4, the majority (81.6 %; 31/38) were clinically insignificant pain symptoms that were incorrectly assessed as clinically significant by ChatGPT GPT-4

* Corresponding author at: University of California, San Francisco, Department of Radiology and Biomedical Imaging, 505 Parnassus Ave, San Francisco, CA 94143, United States of America.

E-mail address: Maggie.chung@ucsf.edu (M. Chung).

https://doi.org/10.1016/j.clinimag.2025.110525
Received 2 December 2024; Received in revised form 19 May 2025; Accepted 22 May 2025
Available online 30 May 2025

Table 1
Summary of clinical classification of breast symptoms by ChatGPT GPT-4: 150 breast pain vignettes. Ground truth clinical significance was determined by the consensus of two fellowship-trained breast radiologists and one breast radiology fellow. True positive (TP) 57, False negative (FN) 7, False positive (FP) 31, True negative (TN) 55.

	Features of breast pain	Ground truth clinical significance	Correct LLM classification of clinical significance [%]	100 % consistency in LLM-assigned clinical significance across three tests [%]
1.	Non-focal (*n* = 15)	Not significant	66.7 (10/15)	80.0 (12/15)
2.	Intermittent/ Cyclical (n = 15)	Not significant	100.0 (15/15)	100.0 (15/15)
3.	Focal/ Intermittent (n = 15)	Not significant	80.0 (12/15)	80.0 (12/15)
4.	Nonfocal/ Constant (n = 15)	Not significant	6.7 (1/15)	86.7 (13/15)
5.	Nonfocal/ Intermittent (n = 15)	Not significant	100.0 (15/15)	93.3 (14/15)
6.	Focal (n = 15)	Significant	53.3 (8/15)	66.7 (10/15)
7.	Focal/Constant (n = 15)	Significant	100.0 (15/15)	100.0 (15/15)
8.	Significant pain + concerning symptom (n = 15)	Significant	100.0 (15/15)	100.0 (15/15)
9.	Nonsignificant pain + concerning symptom (n = 15)	Significant	100.0 (15/15)	100.0 (15/15)
10.	Constant (n = 15)	Mixed significance[a]	40.0 (6/15)	86.7 (13/15)
	All clinically significant symptoms		89.1 (57/64)	92.2 (59/64)
	All clinically insignificant symptoms		64.0 (55/86)	87.2 (75/86)
	All		74.7 (112/150)	89.3 (134/150)

[a] Refers to variation in the clinical characterization of breast symptoms with constant pain.

Table 2
Summary of breast pain cases where LLM-assigned clinical significance differed from breast imaging radiologist consensus by two fellowship-trained breast radiologists and one breast radiology fellow. Classification of LLM-clinical decision making reflects whether the model output represented a clinical upgrade, for example a not significant clinical symptom that GPT-4 considered significant.

	Breast symptom	Clinical feature of pain	Ground truth	LLM-clinical significance	LLM upgrade or downgrade clinical significance
1.	It feels like there's pressure in my entire right breast.	Non-focal	Not significant	Significant	Upgrade
2.	I have pain in my left breast that also goes to the armpit and shoulder.	Non-focal	Not significant	Significant	Upgrade
3.	The pain is spread out across my whole left breast and is even in my armpit.	Non-focal	Not significant	Significant	Upgrade
4.	I have a shooting pain that goes from my left breast to my shoulder.	Non-focal	Not significant	Significant	Upgrade
5.	I have pain that involves my entire right upper outer to upper inner breast.	Non-focal	Not significant	Significant	Upgrade
6.	The soreness in the center of my left breast comes in waves.	Focal/ Intermittent	Not significant	Significant	Upgrade
7.	I occasionally get an intense throbbing pain right in the upper outer part of my right breast. It comes in bursts and then subsides.	Focal/ Intermittent	Not significant	Significant	Upgrade
8.	I feel a sharp, stabbing pain below my right nipple. The pain comes and goes.	Focal/ Intermittent	Not significant	Significant	Upgrade
9.	For the last few days, I've been dealing with this persistent, diffuse pressure in my whole right breast.	Nonfocal/ Constant	Not significant	Significant	Upgrade
10.	I've had a non-stop throbbing pain in my breasts. It's	Nonfocal/ Constant	Not significant	Significant	Upgrade

(continued on next page)

(i.e. upgraded) (Table 2). Of these 31 cases, 5 involved non-focal pain with unspecified temporality (constant vs. intermittent/cyclical), 3 were focal and intermittent pain, 14 were non-focal and constant pain, and 9 were constant pain with unspecified focality. All seven cases with clinically significant pain that were misclassified as insignificant (i.e. downgraded) involved focal pain symptoms without a specified temporality (constant vs. intermittent/cyclical) (Table 2). Upon examination of consistency, 89.3 % (134/150) of ChatGPT GPT-4-generated results were identical in three independent tests.

We demonstrate the first known proposed application of an LLM to automate the classification of patient-centered breast pain symptoms based on clinical significance, which determines the need for diagnostic imaging evaluation. ChatGPT GPT-4 assessments of breast pain symptoms for clinical significance agreed with fellowship-trained breast imaging radiologist consensus in 74.7 % (112/150) of cases, highlighting its potential to streamline patient triaging in busy breast imaging clinics. ChatGPT GPT-4 demonstrated high sensitivity of 89.1 % in identifying clinically significant symptoms. Notably, ChatGPT GPT-4 accurately identified all cases of concerning symptoms associated with either clinically significant or clinically insignificant pain warranting diagnostic evaluation (100 %; 30/30). When ChatGPT GPT-4 made incorrect assessments, it tended to overstate clinical significance. Among the cases where the model disagreed with the radiologists, the majority (81.6 %;

Table 2 (*continued*)

	Breast symptom	Clinical feature of pain	Ground truth	LLM-clinical significance	LLM upgrade or downgrade clinical significance
	been bothering me day and night.				
11.	My breasts have been constantly sore and tender to the touch. The pain spreads all out, and I can't figure out how to make it better.	Nonfocal/ Constant	Not significant	Significant	Upgrade
12.	There's a nagging pain I can feel in my whole left breast.	Nonfocal/ Constant	Not significant	Significant	Upgrade
13.	My whole right breast feels sore and achy all the time.	Nonfocal/ Constant	Not significant	Significant	Upgrade
14.	Both of my breasts hurt throughout the day and night. The pain doesn't go away.	Nonfocal/ Constant	Not significant	Significant	Upgrade
15.	There's pain throughout my left breast that started two weeks ago and hasn't gone away.	Nonfocal/ Constant	Not significant	Significant	Upgrade
16.	I have constant pain throughout my right breast.	Nonfocal/ Constant	Not significant	Significant	Upgrade
17.	The upper and lower parts of my right breast hurt all the time. It doesn't matter what I do.	Nonfocal/ Constant	Not significant	Significant	Upgrade
18.	Both of my breasts ache all the time and the ache never goes away.	Nonfocal/ Constant	Not significant	Significant	Upgrade
19.	I have 3 different spots that hurt all over my upper and lower right breast and 2 spots that hurt in my left breast. The pain seems to be persistent and doesn't go away.	Nonfocal/ Constant	Not significant	Significant	Upgrade
20.	I have pain that involves the entire upper and lower outer right breast. I always notice it.	Nonfocal/ Constant	Not significant	Significant	Upgrade
21.	I have continuous stabbing pain all over my right breast.	Nonfocal/ Constant	Not significant	Significant	Upgrade
22.	There is a nonstop pain that goes from my left arm pit to the outside of my left breast.	Nonfocal/ Constant	Not significant	Significant	Upgrade
23.	I have a sharp pain in one spot on the outside of my right breast	Focal	Significant	Not significant	Downgrade
24.	For the last year, I've had a soreness in one part of my left breast. It gets worse after weightlifting.	Focal	Significant	Not significant	Downgrade
25.	A few days ago, I started feeling a burning sensation at the bottom part of my left breast where my bra usually goes.	Focal	Significant	Not significant	Downgrade
26.	It hurts when I touch a spot at the top of my left breast.	Focal	Significant	Not significant	Downgrade
27.	I have pain in the lower outer part of my right breast that started suddenly a few days ago.	Focal	Significant	Not significant	Downgrade
28.	I have a dull pain in the lower inner right breast.	Focal	Significant	Not significant	Downgrade
29.	I have pain in my left upper outer breast. I have no other symptoms.	Focal	Significant	Not significant	Downgrade
30.	The underside of my right breast hurts all the time.	Constant	Not significant	Significant	Upgrade
31.	I noticed an unrelenting soreness in my left breast for the last two months.	Constant	Not significant	Significant	Upgrade

(continued on next page)

Table 2 (*continued*)

	Breast symptom	Clinical feature of pain	Ground truth	LLM-clinical significance	LLM upgrade or downgrade clinical significance
32.	There's this sharp burning pain along the outside of my right breast.	Constant	Not significant	Significant	Upgrade
33.	I have pain in my right breast that doesn't go away.	Constant	Not significant	Significant	Upgrade
34.	The pain in my right breast feels like a constant pressure or ache.	Constant	Not significant	Significant	Upgrade
35.	My right breast hurts all the time.	Constant	Not significant	Significant	Upgrade
36.	My left breast seems to hurt all the time no matter what kind of bra I wear or what activity I am doing.	Constant	Not significant	Significant	Upgrade
37.	I have noticed a soreness in my left breast for the last 2 weeks that is persistent.	Constant	Not significant	Significant	Upgrade
38.	My left breast is always hurting no matter what I do.	Constant	Not significant	Significant	Upgrade

31/38) involved clinically insignificant pain symptoms that ChatGPT GPT-4 incorrectly classified as clinically significant.

All seven cases of clinically significant pain that were misclassified as insignificant involved focal pain symptoms without a specified temporality (constant vs. intermittent/cyclical). This suggests that ChatGPT GPT-4 may rely on temporality as a key feature when determining clinical significance. Similarly, when assessing cases of clinically insignificant pain, ChatGPT GPT-4 was more likely to misclassify cases when either temporality or focality were not specified, representing nearly half (14/31; 45.2 %) of clinically insignificant cases that were misclassified as significant. Among these, five involved non-focal pain with unspecified temporality and nine cases of constant pain with unspecified focality. ChatGPT GPT-4 also misclassified the majority (14/15; 93.3 %) of cases with non-focal and constant pain. To improve ChatGPT GPT-4 performance, incorporating additional rules in ChatGPT GPT-4 prompt, such as explicitly instructing the model to consider focal pain as significant when temporality is not specified, could help address these gaps. Additionally, further targeted training on these types of cases could further improve ChatGPT GPT-4's performance.

Breast imaging practices, including the author M.C.'s institution, have implemented strategies to reduce unnecessary diagnostic imaging for clinically insignificant pain. These strategies include nursing staff identifying patients with clinically insignificant pain from imaging orders and/or electronic medical record, contacting them to verify symptoms, and rescheduling their unnecessary diagnostic imaging appointments as age-appropriate screening. However, this strategy requires manually reviewing all diagnostic appointments for evaluation of breast pain symptoms, which is time-consuming and burdensome. This approach would be more efficient if breast imaging centers could minimize the number of patients with clinically significant pain in their manual review, as these patients are already scheduled for appropriate diagnostic imaging. Our findings support the potential use of ChatGPT GPT-4 for automated classification of breast pain based on clinical significance, which could potentially be performed by a member of the radiology scheduling team with oversight by a clinical nurse navigator. In this context, the high sensitivity of ChatGPT GPT-4 for clinically significant pain could help rule out patients likely to have clinically significant pain. This would enable breast imaging centers to forgo manual review of 89.1 % of patients with clinically significant symptoms. ChatGPT GPT-4 identified 64.0 % of patients with clinically insignificant pain for whom diagnostic imaging could be safely deferred in favor of age-appropriate screening, in accordance with the ACR Appropriateness Criteria. In the 10.1 % of patients with clinically significant symptoms misclassified as clinically insignificant (false negative), these cases would still undergo review and may require contact. However, this represents a small subset of patients who would have otherwise been needlessly reviewed. Overall, this application has the potential to reduce the need for time-consuming manual review of all breast pain symptoms, allowing focused review and contact of patients most likely to have clinically insignificant pain.

Our findings using patient-centered clinical vignettes of breast pain are aligned with those of a prior study by Rao et al., which found that ChatGPT GPT-3.5 and GPT-4 appropriately recommend screening and diagnostic imaging when they evaluated 3 breast cancer screening and 4 breast pain variants from the respective ACR Appropriateness Criteria.[5] Rao et al. reported that ChatGPT performed with higher accuracy for prompts related to breast cancer screening compared to breast pain, reporting >50 % accuracy in breast pain cases ($n = 4$). However, their study relied on clinical assessments of breast pain, such as focal and nonfocal pain in prompting. This approach limits applicability in real-world settings, as it does not account for the diverse ways in which patients describe their pain. In this study, we simulate patient-reported descriptors of pain and use ChatGPT GPT-4 to classify these descriptions, better reflecting real-world scenarios and potential applications. This complex task requires an assessment of layperson descriptions of breast pain and a classification of the clinical significance of breast pain symptoms. Although the clinical history provided in imaging study orders is often nonspecific, such as "pain," future applications of ChatGPT GPT-4 may be able to extract and summarize this information from elsewhere in the electronic health record. ChatGPT GPT-4 performed in alignment with the breast imaging radiologists in nearly three-quarters of cases, encompassing the clinical variants described in the ACR Appropriateness Criteria.

This study has limitations. Pain symptoms were simulated and written by experienced breast radiologists to reflect how patients might commonly describe their symptoms, rather than pre- populated study indication frequently used by referring clinicians in the electronic health record. Future studies should evaluate the model prospectively using patient-provided symptoms to better assess the potential performance in real-world scenarios with a larger sample size. The study evaluated the performance of zero-shot ChatGPT GPT-4 prompting. Additional studies are needed to evaluate the performance of other LLMs, as well as the potential benefits of fine-tuning and prompt engineering, to further enhance the model's accuracy. Inconsistency among the breast pain classifications is likely attributable to the probabilistic nature of LLM, whereby responses may vary even when the input is unchanged.

In conclusion, LLMs offer a distinct opportunity to efficiently classify patient reported pain symptoms to help inform the appropriate imaging needs based on the ACR's recommendations and improve imaging resource utilization, though current performance with a non-trivial downgrade rate precludes autonomous clinical use in the current form.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinimag.2025.110525.

CRediT authorship contribution statement

Hana Haver: Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization, Writing – review & editing, Writing – original draft. **Manisha Bahl:** Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization, Writing – review & editing, Writing – original draft. **Maggie Chung:** Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization, Writing – review & editing, Writing – original draft.

Declaration of competing interest

The authors declare that the work described has not been published previously except in the form of a preprint, an abstract, a published lecture, academic thesis or registered report. This work was presented as a scientific poster at the 2024 Conference on Machine Intelligence in Medical Imaging Conference.

The article is not under consideration for publication elsewhere.

The article's publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out.

If accepted, the article will not be published elsewhere in the same form, in English or in any other language, including electronically, without the written consent of the copyright-holder.

References

1. Holbrook AI, Moy L, Akin EA. ACR Appropriateness Criteria® breast pain [internet]. American College of Radiology. Available from: https://acsearch.acr.org/docs/3091546/Narrative/.
2 Chetlen AL, Kapoor MM, Watts MR. Mastalgia: imaging work-up appropriateness. Acad Radiol 2017;24:345–9.
3 Olcucuoglu E, Yilmaz G. Mastodynia: is imaging necessary in young patients? Ulus Cerrahi Derg 2013;29:17–9 [36. Harper AP, Kel].
4 Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology 2023 Apr 4;307(4):230424.
5 Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 2023 Oct;20(10):990–7. S1546144023003940.