3° Πανελλήνιο Συνέδριο
Ιατρικής Φυσικής

27° Διαπανεπιστημιακό Συνέδριο
Ακτινολογίας

Πόσο αξιόπιστη είναι η τεχνητή νοημοσύνη στη διάγνωση της οξείας κοιλίας με αξονική τομογραφία;

Συγκριτική αποτίμηση τριών γλωσσικών μοντέλων αιχμής

Κουρούπη Αλεξάνδρα, Σαριδάκη Μαρία, Μάνος Βάιος,
Σύψα Γεωργία, Βεργαδής Χρυσοβαλάντης, Κελέκης Αλέξιος-Χρυσόστομος

Γ΄ Εργαστήριο Ακτινολογίας, ΕΚΠΑ, ΓΝΑ Λαϊκό, Αθήνα

30, 31 ΟΚΤ + 1 ΝΟΕ
2026

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
- ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ, ΛΑΡΙΣΑ

Οργάνωση

ΕΦΙΕ
Ένωση Φυσικών Ιατρικής Ελλάδος

ΕΛΛΗΝΙΚΟ ΚΟΛΛΕΓΙΟ ΑΚΤΙΝΟΛΟΓΙΑΣ

Εισαγωγή

Κλινικό πλαίσιο

Το οξύ κοιλιακό άλγος αποτελεί συχνή αιτία προσέλευσης στα επείγοντα, με την αξονική τομογραφία (CT) να διαδραματίζει κεντρικό ρόλο στην έγκαιρη διάγνωση επειγουσών παθήσεων. Τα μεγάλα γλωσσικά μοντέλα (Large Language Models, LLMs) διερευνώνται ως πιθανά εργαλεία υποστήριξης της ακτινολογικής ερμηνείας. [1,2]

Τι αξιολογήθηκε

Αξιολογήθηκαν η κύρια διάγνωση, η κατάταξη της σωστής διάγνωσης στη διαφορική διάγνωση, η αναγνώριση των αληθών απεικονιστικών ευρημάτων, η δηλωμένη βεβαιότητα, τα μη τεκμηριωμένα ευρήματα και η αιτιολόγηση. Η συνολική επίδοση συνοψίστηκε σε σύνθετη βαθμολογία.

Σκοπός

Συγκριτική αξιολόγηση της διαγνωστικής απόδοσης τριών γλωσσικών μοντέλων αιχμής - Claude Opus 4.7, GPT-5.5 και Gemini 3.1 Pro - σε 100 CT κοιλίας από το Τμήμα Επειγόντων Περιστατικών, με πρότυπο αναφοράς τις ακτινολογικές εκθέσεις.

Φάσμα διαγνώσεων

Σκωληκοειδίτιδα, εκκολπωματίτιδα, ειλεός, διάτρηση, χολοκυστίτιδα, παγκρεατίτιδα, ουρολιθίαση, αιμορραγία, απόστημα, ισχαιμία, κήλη και νεόπλασμα.

Μέθοδοι & Υλικά

Σχεδιασμός και δείγμα

Αναδρομική ανάλυση 100 τυχαία επιλεγμένων, ανωνυμοποιημένων CT κοιλίας ενηλίκων από αρχεία Τμήματος Επειγόντων Περιστατικών.

Δεδομένα και πρότυπο αναφοράς

Για κάθε περιστατικό δόθηκαν στα τρία μοντέλα τα ίδια δεδομένα εισόδου. Η τελική ακτινολογική έκθεση χρησιμοποιήθηκε ως πρότυπο αναφοράς για τη σύγκριση των απαντήσεων.

Προκαθορισμένη μορφή απάντησης

Κάθε μοντέλο παρείχε κύρια διάγνωση, βεβαιότητα 0-100, 2-5 διαφορικές διαγνώσεις κατά σειρά πιθανότητας, κύρια και παρεμπίπτοντα ευρήματα, καθώς και σύντομη αιτιολόγηση.

Τυφλοποίηση και αξιολόγηση

Οι απαντήσεις κωδικοποιήθηκαν ως LLM1/LLM2/LLM3. Το Claude Opus 4.7 τις συνέκρινε με την έκθεση αναφοράς βάσει προκαθορισμένων κριτηρίων.

Σύνθετη βαθμολογία

Συνιστώσα	Βάρος
Διαγνωστική ακρίβεια	50%
Κατάταξη διαφορικής διάγνωσης	15%
Αναγνώριση αληθών απεικονιστικών ευρημάτων	15%
Βαθμονόμηση βεβαιότητας	10%
Αιτιολόγηση	10%

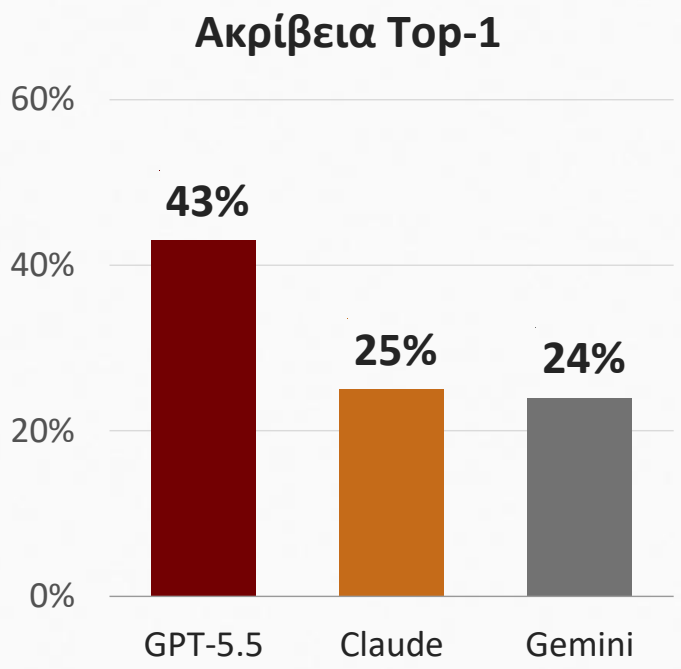
Στατιστική ανάλυση

Σύνθετη βαθμολογία: ζευγαρωτές δοκιμασίες προσήμου (sign tests).

Top-1 ακρίβεια: 95% ΔΕ κατά Wilson.

Αποτελέσματα

Δείκτης	GPT-5.5	Claude	Gemini
Ακρίβεια Top-1, % (95% ΔΕ)	43 (34-53)	25 (18-34)	24 (17-33)
Ακρίβεια Top-5, %	65	51	46
Μέση αντίστροφη κατάταξη (MRR)	0,52	0,35	0,34
Μέση σύνθετη βαθμολογία	63,7	49,5	47,7
Πρώτη θέση, περιστατικά / 100	56	31	13
Μη τεκμηριωμένα ευρήματα / περιστατικό	1,4	1,9	2,1
Top-1 σε δύσκολα περιστατικά, %[1]	9	0	5



Σύγκριση σύνθετης βαθμολογίας: ζευγαρωτές δοκιμασίες προσήμου (sign tests)

GPT-5.5 > Claude: p = 0,006. GPT-5.5 > Gemini: p < 0,001. Claude > Gemini: p < 0,001.

Top-1: σωστή κύρια διάγνωση. Top-5: σωστή διάγνωση στις πρώτες πέντε θέσεις. ΔΕ: διάστημα εμπιστοσύνης. MRR: μέση αντίστροφη θέση της σωστής διάγνωσης (υψηλότερη = καλύτερη). [1] Δύσκολα περιστατικά: n = 22.

Συμπεράσματα

Διαγνωστικές αποτυχίες

Και τα τρία μοντέλα έδωσαν σωστή κύρια διάγνωση μόνο σε 11/100 περιστατικά. Σε 27/100, η σωστή διάγνωση δεν εμφανίστηκε σε καμία απάντηση. Στα 22 δύσκολα περιστατικά, η Top-1 ακρίβεια ήταν 9% για το GPT-5.5, 0% για το Claude και 5% για το Gemini.

Ερμηνεία των ευρημάτων

Στο συγκεκριμένο δείγμα και με την παρούσα μεθοδολογία, το GPT-5.5 είχε την καλύτερη συνολική επίδοση, χωρίς όμως να φτάνει επίπεδο ασφαλούς αυτόνομης διάγνωσης. Τα αποτελέσματα υποστηρίζουν πιθανή χρήση ως εργαλείο υποστήριξης ή δεύτερης γνώμης, όχι ως ανεξάρτητο διαγνώστη CT.

Βεβαιότητα και ακρίβεια

Η υψηλή δηλωμένη βεβαιότητα δεν σήμαινε απαραίτητα σωστή διάγνωση. Το Gemini δήλωνε μέση βεβαιότητα 94,6/100, αλλά είχε Top-1 ακρίβεια 24% και πολύ μικρή διαφορά βεβαιότητας μεταξύ σωστών και λανθασμένων απαντήσεων (1,6 μονάδες).

Περιορισμοί και γενίκευση

Η μελέτη περιέλαβε περιορισμένο αριθμό περιστατικών από ένα μόνο κέντρο και δεν πραγματοποιήθηκε ανεξάρτητη εξωτερική επικύρωση. Συνεπώς, τα ευρήματα χρειάζονται επιβεβαίωση σε μεγαλύτερα και ανεξάρτητα δείγματα ασθενών.

Βιβλιογραφία

1. Mukherjee P, Hou B, Suri A, et al. Evaluation of GPT Large Language Model Performance on RSNA 2023 Case of the Day Questions. Radiology. 2024;313(1):e240609.

   doi: 10.1148/radiol.240609

2. 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):2613-2622.

   doi: 10.1038/s41591-024-03097-1

3. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Adv Neural Inf Process Syst. 2023;36:46595-46623.

   doi: 10.52202/075280-2020

Σύγκρουση συμφερόντων: καμία, σύμφωνα με τη δήλωση της υποβολής.