Αξιολόγηση της Αξιοπιστίας Συστημάτων Τεχνητής Νοημοσύνης στην Κριτική Επιστημονικών Περιλήψεων Υποβληθέντων στο Πανελλήνιο Καρδιολογικό Συνέδριο

ΑΝΑΣΤΑΣΙΟΣ ΑΠΟΣΤΟΛΟΣ[1,2], ΙΩΑΝΝΗΣ ΚΑΧΡΙΜΑΝΙΔΗΣ[1], ΝΙΚΟΛΑΟΣ ΚΤΕΝΟΠΟΥΛΟΣ[1], ΝΙΚΟΛΑΟΣ ΠΑΤΣΟΥΡΑΚΟΣ[3], ΑΓΓΕΛΙΚΗ-ΔΕΣΠΟΙΝΑ ΜΑΥΡΟΓΙΑΝΝΗ[4], ΗΛΙΑΣ ΚΑΡΑΜΠΙΝΟΣ[5], ΑΡΗΣ ΑΝΑΣΤΑΣΑΚΗΣ[6], ΛΟΥΚΙΑΝΟΣ ΡΑΛΛΙΔΗΣ[7], ΑΘΑΝΑΣΙΟΣ ΤΡΙΚΑΣ[8], ΓΕΩΡΓΙΟΣ ΚΟΧΙΑΔΑΚΗΣ[9], ΛΑΜΠΡΟΣ ΜΙΧΑΛΗΣ[10], ΚΩΝΣΤΑΝΤΙΝΟΣ ΤΟΥΤΟΥΖΑΣ[1]

[1] Α' Καρδιολογική Κλινική, Ιατρική Σχολή Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών, ΓΝΑ Ιπποκράτειο
[2] Department of Cardiology, Harefield Hospital, Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom
[3] Καρδιολογικό τμήμα, Γενικό Νοσοκομείο Πειραιά "Τζάνειο"
[4] Καρδιολογική Κλινική, Γενικό Νοσοκομείο Θεσσαλονίκης "Γ. Παπανικολάου"
[5] Β' Καρδιολογική Κλινική, Ευρωκλινική Αθηνών
[6] Ωνάσειο Καρδιοχειρουργικό Κέντρο
[7] Β' Καρδιολογική Κλινική, Ιατρική Σχολή Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών, ΠΓΝΑ Αττικόν
[8] Καρδιολογικό Τμήμα, ΓΝΑ Ευαγγελισμός
[9] Καρδιολογική Κλινική, ΠΑΓΝΗ
[10] Καρδιολογική Κλινική, ΠΓΝΙ

Λέξεις ευρετηρίου
Τεχνητή νοημοσύνη, ChatGPT, DeepSeek, Grok, αξιολόγηση περιλήψεων συνεδρίου, προοπτική μελέτη

Επικοινωνία
Αναστάσιος Αποστολός, MD, MSc
Υποψήφιος Διδάκτωρ, Α' Καρδιολογική Κλινική, Ιατρική Σχολή, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, ΓΝΑ «Ιπποκράτειο»
E-mail: anastasisapostolos@gmail.com

Η συνεχής αύξηση του αριθμού των περιλήψεων που υποβάλλονται σε ιατρικά συνέδρια, και ειδικότερα στο Πανελλήνιο Καρδιολογικό Συνέδριο, καθιστά την αξιολόγησή τους ολοένα και πιο απαιτητική, τόσο από πλευράς χρόνου όσο και από πλευράς διασφάλισης αντικειμενικότητας. Η πρόσφατη ανάπτυξη μεγάλων γλωσσικών μοντέλων, όπως το ChatGPT, το DeepSeek και το Grok, προσφέρει νέες δυνατότητες στον χώρο της επεξεργασίας επιστημονικού λόγου και θέτει το ερώτημα κατά πόσο μπορούν να χρησιμοποιηθούν ως επικουρικά εργαλεία στη διαδικασία της επιστημονικής κρίσης. Στόχος της παρούσας μελέτης είναι η συγκριτική αποτίμηση της συμφωνίας μεταξύ ανθρώπινων αξιολογητών και συστημάτων τεχνητής νοημοσύνης στην κρίση ανωνυμοποιημένων περιλήψεων που θα υποβληθούν στο Πανελλήνιο Καρδιολογικό Συνέδριο 2025. Η μελέτη έχει προοπτικό χαρακτήρα και προβλέπει τη διπλή αξιολόγηση των περιλήψεων από ανθρώπινους κριτές και από τρία διαφορετικά μεγάλα γλωσσικά μοντέλα, με χρήση τυποποιημένων προτροπών. Σε περίπτωση υψηλής συμφωνίας, η μελέτη θα μπορούσε να αποτελέσει το πρώτο βήμα για την υιοθέτηση ενός υβριδικού μοντέλου κρίσης, ενώ σε περίπτωση χαμηλής συμφωνίας θα αναδείξει τους κινδύνους και τα όρια της τεχνολογίας.

Εισαγωγή

Η συνεχώς αυξανόμενη παραγωγή επιστημονικής γνώσης στον χώρο της ιατρικής έχει ως αποτέλεσμα την εντυπωσιακή αύξηση των περιλήψεων που υποβάλλονται σε συνέδρια. Τα εθνικά και διεθνή καρδιολογικά συνέδρια δέχονται πλέον χιλιάδες υποβολές ετησίως, γεγονός που δημιουργεί σημαντικές προκλήσεις ως προς την έγκαιρη, αξιόπιστη και αντικειμενική αξιολόγηση των εργασιών. Το παραδοσιακό σύστημα κριτών (peer-review) παραμένει ο θεμέλιος λίθος της επιστημονικής αξιολόγησης, εντούτοις η διαδικασία του είναι χρονοβόρα, εξαρτάται

από την εμπειρία και τη διαθεσιμότητα των αξιολογητών, ενώ δεν στερείται υποκειμενικότητας.

Τα τελευταία έτη, η ανάπτυξη μεγάλων γλωσσικών μοντέλων (Large Language Models, LLMs), όπως το ChatGPT (OpenAI), το DeepSeek (China-based) και το Grok (X), προσέφερε νέες δυνατότητες στον τομέα της επεξεργασίας κειμένου και της αυτόματης γλωσσικής ανάλυσης.[1–6] Οι εφαρμογές τους στον χώρο της ιατρικής εκτείνονται από την εκπαίδευση έως την υποστήριξη κλινικών αποφάσεων. Ιδιαίτερο ενδιαφέρον παρουσιάζει η πιθανή αξιοποίησή τους στην επιστημονική κριτική (peer review), με στόχο τη βελτίωση της αποδοτικότητας και της αντικειμενικότητας της διαδικασίας.

Ωστόσο, η αξιοπιστία των συστημάτων αυτών παραμένει αντικείμενο υπό διερεύνηση. Πρόσφατες μελέτες έχουν αναδείξει ότι τα LLMs μπορούν να προσφέρουν εποικοδομητικά σχόλια συγκρίσιμα με αυτά ανθρώπινων κριτών, τουλάχιστον σε ορισμένα πλαίσια. Για παράδειγμα, σε μελέτη της Liang και συνεργατών, διαπιστώθηκε ότι η επικάλυψη σχολίων GPT-4 με ανθρώπινους κριτές σε περιοδικά της οικογένειας Nature ήταν παρόμοια με την επικάλυψη μεταξύ δύο ανθρώπων κριτών.[7] Παράλληλα, μελέτη των Saad και συνεργατών έδειξε ότι οι αξιολογήσεις του ChatGPT συσχετίζονται σε σημαντικό βαθμό με εκείνες των ανθρώπων.[8]

Παρά τις ενθαρρυντικές αυτές ενδείξεις, η αξιοπιστία και η καταλληλόλητα των LLMs για χρήση σε καρδιολογικά συνέδρια δεν έχει ακόμη μελετηθεί επαρκώς. Η παρούσα μελέτη φιλοδοξεί να καλύψει το κενό αυτό, αξιολογώντας συστηματικά τρία διαφορετικά συστήματα, με βάση τις περιλήψεις του Πανελλήνιου Καρδιολογικού Συνεδρίου 2025.

Μεθοδολογία

Η μελέτη έχει προοπτικό και συγκριτικό χαρακτήρα. Ο πληθυσμός της μελέτης θα περιλαμβάνει όλες τις περιλήψεις που θα υποβληθούν στο Πανελλήνιο Καρδιολογικό Συνέδριο 2025. Οι περιλήψεις θα ανωνυμοποιηθούν πλήρως, ώστε να αφαιρεθούν όλα τα στοιχεία που θα μπορούσαν να αποκαλύψουν την ταυτότητα των συγγραφέων. Κάθε περίληψη θα αξιολογηθεί μέσω δύο παράλληλων διαδικασιών. Στην πρώτη, οι κριτές του συνεδρίου θα εξετάσουν τις περιλήψεις σύμφωνα με το υφιστάμενο σύστημα, βαθμολογώντας τες σε κλίμακα 1 έως 10 και αποφασίζοντας για την αποδοχή ή την απόρριψή τους. Στη δεύτερη, οι ίδιες περιλήψεις θα εισαχθούν σε τρία LLMs (ChatGPT, DeepSeek, Grok) με χρήση τυποποιημένων εντολών. Τα συστήματα θα κληθούν να παραγάγουν σύντομα θετικά και αρνητικά σχόλια και να αποδώσουν βαθμολογία 1 έως 10, όπως αντίστοιχα και οι ανθρώπινοι κριτές. Η στατιστική ανάλυση θα εστιάσει στον υπολογισμό του συντελεστή Cohen's kappa για τη συμφωνία μεταξύ ανθρώπων και LLMs, καθώς και μεταξύ διαφορετικών LLMs. Επιπλέον, θα πραγματοποιηθεί ROC curve analysis για την αποτίμηση της προγνωστικής ικανότητας των μοντέλων ως προς την τελική αποδοχή, ενώ οι ποσοτικές αξιολογήσεις θα συγκριθούν μέσω Bland-Altman plots. Επιπλέον, αναλύσεις θα διεξαχθούν με βάση τη θεματολογία της περίληψης, τον τύπο της μελέτης και το είδος του LLM. Όλα τα δεδομένα θα είναι πλήρως ανωνυμοποιημένα, χωρίς συλλογή ή αποθήκευση προσωπικών δεδομένων. Η μελέτη έχει λάβει έγκριση από το διοικητικό συμβούλιο της Ελληνικής Καρδιολογικής Εταιρείας.

Συζήτηση

Η μελέτη αναμένεται να αξιολογήσει και να εκτιμήσει τον βαθμό στον οποίο τα LLMs μπορούν να αναπαράγουν ή να προσεγγίσουν την κρίση των ανθρώπινων αξιολογητών στις περιλήψεις που υποβάλλονται σε καρδιολογικά συνέδρια. Εάν τα αποτελέσματα δείξουν υψηλή συμφωνία, θα ανοίξει ο δρόμος για τη χρήση των συστημάτων αυτών ως επικουρικών εργαλείων στη διαδικασία κρίσης, είτε για προεπιλογή περιλήψεων είτε για υποστήριξη των κριτών. Αυτό θα μπορούσε να οδηγήσει σε ταχύτερη, πιο αντικειμενική και πιο αποδοτική διαδικασία αξιολόγησης. Αντιστρόφως, χαμηλή συμφωνία θα καταδείξει ότι οι δυνατότητες των LLMs είναι ακόμη περιορισμένες και ότι απαιτείται περαιτέρω βελτίωση πριν από οποιαδήποτε ενσωμάτωση σε πραγματικές συνθήκες. Η διεθνής βιβλιογραφία καταδεικνύει ότι η χρήση τεχνητής νοημοσύνης στο peer review βρίσκεται σε πρώιμο, αλλά ραγδαία εξελισσόμενο στάδιο. Με-

λέτες όπως εκείνες των Li και συνεργατών στο JAMA Network Open και των Liang και συνεργατών στο NEJM AI προσέφεραν ισχυρές ενδείξεις ότι τα LLMs μπορούν να παράσχουν σχόλια συγκρίσιμα με εκείνα των ανθρώπων.[7,9] Η δική μας μελέτη διαφοροποιείται καθώς εφαρμόζεται σε αξιολόγηση περιλήψεων συνεδρίου και δεν αφορά πλήρη κείμενα με πραγματικά δεδομένα από τον χώρο της καρδιολογίας και στοχεύει να διερευνήσει την κλινική αξία μιας τέτοιας προσέγγισης στο συγκεκριμένο επιστημονικό πλαίσιο. Σε περίπτωση που αναδειχθούν θετικά αποτελέσματα, θα μπορούσε να τεθεί το υπόβαθρο για τη δημιουργία ενός υβριδικού μοντέλου, όπου οι άνθρωποι κριτές και τα LLMs συνεργάζονται συμπληρωματικά. Ωστόσο, ακόμη και σε περίπτωση αρνητικών ευρημάτων, η μελέτη θα έχει αξία καθώς θα υπογραμμίσει τις αδυναμίες των συστημάτων αυτών και θα προσδιορίσει τα όρια ασφαλούς εφαρμογής τους.

Συμπεράσματα

Η αξιολόγηση επιστημονικών περιλήψεων αποτελεί κρίσιμη διαδικασία στη διάχυση της γνώσης και στη διασφάλιση της ποιότητας των συνεδρίων. Η παρούσα μελέτη, η οποία θα διεξαχθεί στο πλαίσιο του Πανελλήνιου Καρδιολογικού Συνεδρίου 2025, φιλοδοξεί να προσφέρει πρωτογενή στοιχεία για την αξιοπιστία των LLMs στη διαδικασία της επιστημονικής κρίσης. Τα αποτελέσματά της θα έχουν σημασία όχι μόνο σε εθνικό, αλλά και σε διεθνές πλαίσιο, συνεισφέροντας στην ευρύτερη συζήτηση γύρω από την ενσωμάτωση της τεχνητής νοημοσύνης στην ιατρική επιστήμη.

Βιβλιογραφία

1. Buess L, Keicher M, Navab N, Maier A, Tayebi Arasteh S. From large language models to multimodal AI: a scoping review on the potential of generative AI in medicine. Biomed Eng Lett. 2025;15(5):845-863.
2. Almufarreh A, Ahmad A, Arshad M, Onn CW, Elechi R. Ethical implications of ChatGPT and other large language models in academia. Front Artif Intell. 2025;8.
3. Pellegrina D, Helmy M. AI for scientific integrity: detecting ethical breaches, errors, and misconduct in manuscripts. Front Artif Intell. 2025;8.
4. Rose CJ, Bidonde J, Ringsten M, et al. Using a Large Language Model (ChatGPT-4o) to Assess the Risk of Bias in Randomized Controlled Trials of Medical Interventions: Interrater Agreement With Human Reviewers. Cochrane Evidence Synthesis and Methods. 2025;3(5).
5. Skalidis I, Cagnina A, Luangphiphat W, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? European Heart Journal - Digital Health. Published online April 24, 2023.
6. Chlorogiannis DD, Apostolos A, Chlorogiannis A, et al. The Role of ChatGPT in the Advancement of Diagnosis, Management, and Prognosis of Cardiovascular and Cerebrovascular Disease. Healthcare. 2023;11(21):2906.
7. Liang W, Zhang Y, Cao H, et al. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. NEJM AI. 2024;1(8).
8. Saad A, Jenko N, Ariyaratne S, et al. Exploring the potential of ChatGPT in the peer review process: An observational study. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2024;18(2):102946.
9. Li ZQ, Xu HL, Cao HJ, Liu ZL, Fei YT, Liu JP. Use of Artificial Intelligence in Peer Review Among Top 100 Medical Journals. JAMA Netw Open. 2024;7(12): e2448609.