

LLM-Assisted Screening for Plastic Surgery Systematic Reviews

Neil Reddy¹, Nihal Sriramaneni¹, Nitya Devisetti¹, Justin Park¹, Pushti Shah¹, Christopher Didzbalis¹, Alex Wong¹

¹Rutgers New Jersey Medical School, Newark, NJ, USA



Division of Plastic and Reconstructive Surgery
Department of Surgery
Center for Inflammation & Immunity

Purpose

Systematic reviews are valuable to plastic surgery researchers, but abstract screening is time-intensive and variable. Large language models (LLMs) such as ChatGPT may accelerate title/abstract and full-text screening, but optimal prompting is unclear. We evaluated LLM performance for screening in plastic surgery systematic reviews and tested whether PICO statements or inclusion/exclusion (I/E) criteria yield better diagnostic performance. We also compared multiple LLMs to identify the most accurate option.

Methods

A random sample of 50 papers drawn from three completed systematic reviews, with human decisions blinded to the LLMs, served as the reference standard. The topics focused on the intersection of plastic surgery and the lymphatic system. We evaluated two screening contexts: abstract and full text. Within each context, prompts supplied either PICO alone, I/E alone, or both. For the combined condition, results from the PICO-only and I/E-only prompts were combined such that a paper was excluded only if both prompts excluded it. Primary outcomes included confusion matrices and derived metrics, including sensitivity, specificity, PPV, NPV, accuracy, and Matthews correlation coefficient. In a separate comparison within one review, we tested ChatGPT 5, SciSpace Agent, and Claude Sonnet 4.5 using PICO alone, I/E alone, and a combined prompting strategy that required agreement among all LLMs for exclusion. Agreement with humans and across LLMs was assessed using Cohen's κ and Fleiss' κ , respectively.

Results

For abstract screening, mean sensitivity was 0.61 ± 0.21 (PICO), 0.70 ± 0.15 (I/E), and 0.85 ± 0.14 (combined). For full-text screening, sensitivity was 0.71 ± 0.25 (PICO), 0.56 ± 0.34 (I/E), and 0.90 ± 0.08 (combined). In the model comparison (combined prompting), sensitivities were: ChatGPT 0.86, SciSpace 0.79, and Claude 0.86. Agreement with human screening (Cohen's κ) was 0.71 (ChatGPT), 0.58 (SciSpace), and 0.53 (Claude). Cross-model agreement (Fleiss' κ) was 0.68; including manual screening yielded $\kappa = 0.64$. Combining the results of all three LLMs increased sensitivity to 0.93.

ChatGPT Combined PICO + I/E Screening Results

	Abstract	Full-Text
Sensitivity	0.85 +/- 0.14	0.90 +/- 0.08
Specificity	0.49 +/- 0.36	0.88 +/- 0.10
PPV	0.35 +/- 0.21	0.66 +/- 0.25
NPV	0.96 +/- 0.03	0.97 +/- 0.03
Accuracy	0.57 +/- 0.27	0.88 +/- 0.08
Matthews Correlation Coefficient	0.33 +/- 0.23	0.70 +/- 0.15

Figure 1: Results of combined PICO + I/E abstract and full-text screening using ChatGPT 5. N=150 papers. Results are reported as Average +/- SD

Combined LLM Screening Results

	PICO	I/E	PICO + I/E
I/E Sens	0.92857143	0.92857143	0.92857143
I/E Spec	0.72222222	0.72222222	0.69444444
I/E PPV	0.56521739	0.56521739	0.54166667
I/E NPV	0.96296296	0.96296296	0.96153846
Accuracy	0.78	0.78	0.76
Matthews Correlation Coefficient	0.58629039	0.58629039	0.55991498

Figure 2: Results of combined LLM full text screening. N=50 Papers.

Results (continued)

Comparison of LLM Screening Results

	Cohen's κ	Accuracy	Sensitivity	Specificity
ChatGPT	0.71	0.88	0.86	0.89
Scispace	0.58	0.82	0.79	0.83
Claude	0.53	0.78	0.86	0.75

Figure 3: Results of combined PICO + I/E full-text screening context for each LLM. N=50 Papers

Discussion and Conclusions

LLM-assisted screening can streamline plastic surgery systematic reviews but should be applied cautiously. In our analysis, sensitivity did not meet the Cochrane automated-screening minimum of 0.99. Across abstract and full-text tasks, using both PICO and I/E criteria—and combining their decisions—consistently improved sensitivity over single-prompt approaches, yielding the most favorable sensitivity–specificity balance in our cohort. Among tested models, ChatGPT-5 and Claude Sonnet 4.5 showed similar sensitivities, with SciSpace lower; ChatGPT-5 also achieved higher specificity, PPV, NPV, accuracy, and Matthews correlation coefficient. Combining LLM decisions produced the highest sensitivity, underscoring the value of multi-method screening for improved study capture. All models reliably filtered clearly irrelevant articles but were less adept at identifying all eligible ones, underscoring the need for human oversight. Further work on prompt design and API-level controls (e.g., temperature and other inference parameters) may improve sensitivity. Prospective studies should test these configurations across additional topics and evaluate usability, integration, and reproducibility.