IFSO

XXIX IFSO World Congress

1-4 September 2026 | Toronto, Canada

Performance of Artificial Intelligence in Metabolic and Bariatric Surgery

Jerry Dang, MD, PhD, FRCSC, FACS, FASMBS

Associate Professor of Surgery, Cleveland Clinic

2 September 2026

TORONTO2026

The Future of Obesity Management:
From Weight Loss to Health Gain

ifso2026.org

Disclosures

Consultant — SmartSleeveMD

Our Journey Today

01

Patient-Forward

AI coaching, chatbots, LLMs

02

Clinician-Forward

LLMs, ambient scribes, appeal letters

03

The Future

International expert consensus

01

Patient-Forward Applications

Can an AI coach match a human coach for weight loss?

Randomized non-inferiority trial · 368 adults with prediabetes + overweight/obesity · fully automated AI app (reinforcement-learning push notifications + Bluetooth scale) vs CDC-recognized human-coached Diabetes Prevention Program (DPP) · 12 months · non-inferiority margin −15%

31.7% vs **31.9%**

met the CDC composite outcome (≥5% weight loss, or ≥4% + activity, or HbA1c drop), AI vs human — non-inferior

16.9% vs **20.0%**

achieved ≥5% weight loss, AI vs human (RD −3.1, CI −9.7); HbA1c reduction ≥0.2 identical, 26.9% each

93% vs **83%**

started the program; completion 64% vs 50% (both p ≤ 0.008)

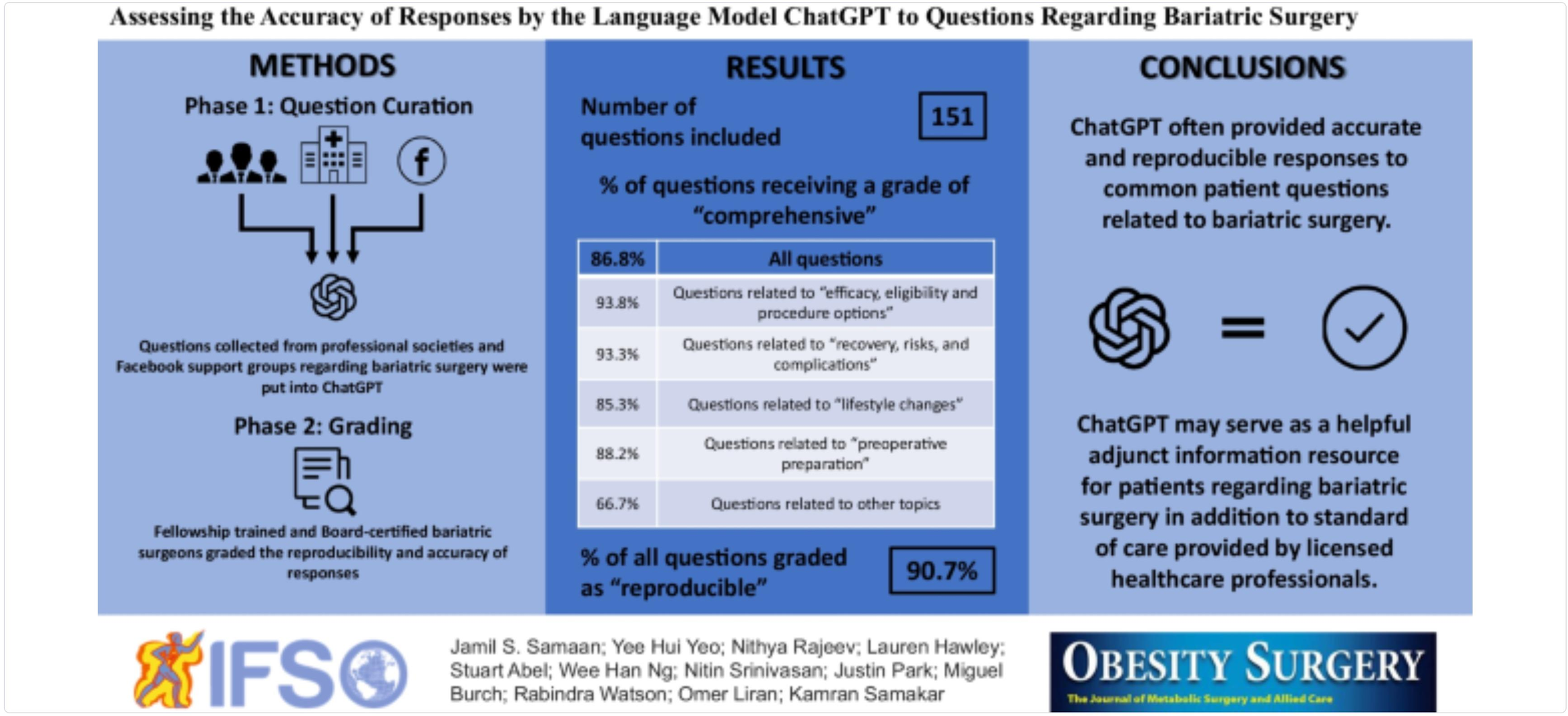
AI delivers the **same modest result** as human coaching

— but **scales** and removes the scheduling barrier

Caveat: motivated, educated volunteers at 2 sites

Mathioudakis N, Lalani B, et al. An AI-Powered Lifestyle Intervention vs Human Coaching in the Diabetes Prevention Program: A Randomized Clinical Trial. JAMA 2025;334(23):2079–2089.

Can LLMs help patients understand bariatric surgery?



Samaan JS, et al. Obes Surg 2023;33:1790–1796. · ChatGPT-3.5, February 2023

LLM vs surgeon answers — who do patients prefer?

200 real patient questions · GPT-4o vs 2 bariatric surgeons · blinded expert review · rated by 189 patients

64.9%
of patients preferred the LLM answer; 18.5% preferred the surgeon

4.1 vs 3.2
empathy, LLM vs surgeon (1–5); completeness 4.5 vs 3.4

2.7 s
to generate an answer vs 87 s for surgeons

6.5%
of LLM answers needed correction — one potentially dangerous

WHAT IT MEANS

LLMs write faster, warmer and more complete patient-facing answers than busy surgeons do

THE CATCH

1 in 15 answers still needed a clinician to fix it — deploy as a drafting tool, not an unsupervised one

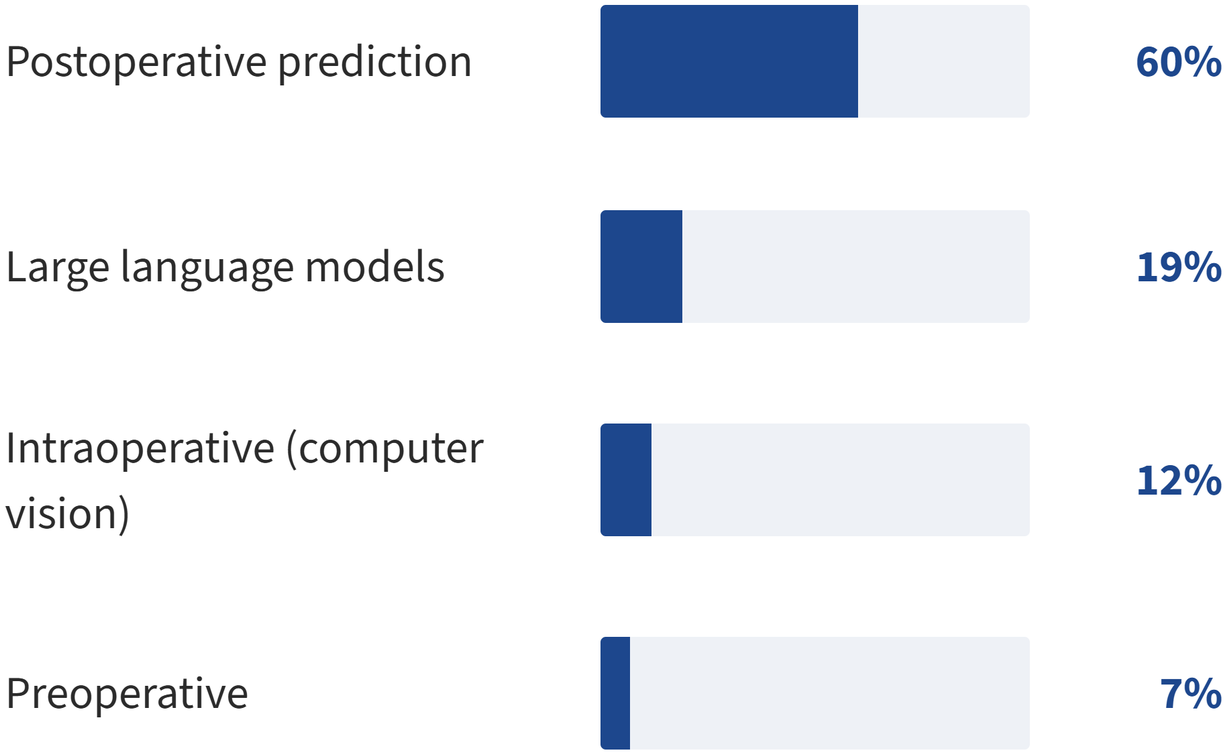
Vedder K, et al. Obes Surg 2025;35:3801–3809.

02

Clinician-Forward Applications

Where is AI being used along the MBS pathway?

Scoping review · 4 databases to Nov 2024 · 58 studies · 55% published in 2023–24



Neural networks in 55% of ML studies · ChatGPT in 10 of 11 LLM studies · MBSAQIP the most common data source

0

prospective studies or randomized trials — all 58 retrospective or cross-sectional

Minimal

external validation; almost no direct clinical evaluation

Brar K, Chattha I, Gobraeil S, Mocanu V, Corcelles R, Kroh M, Shin T, Dang JT, Lee Y. Surg Obes Relat Dis 2026;22:1025–1034.

Can LLMs answer guideline-based clinical questions?

ASMBS Guidelines/Statements

Harnessing artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and Bard in generating clinician-level bariatric surgery recommendations

Yung Lee, M.D., M.PH.[a,b], Thomas Shin, M.D., Ph.D.[c], Léa Tessier, M.D.[a], Arshia Javidan, M.D., M.Sc.[d], James Jung, M.D., Ph.D.[e], Dennis Hong, M.D., M.Sc.[a], Andrew T. Strong, M.D.[f], Tyler McKechnie, M.D., M.Sc.[a], Sarah Malone, B.H.Sc.[a], David Jin, B.H.Sc.[a], Matthew Kroh, M.D.[f], Jerry T. Dang, M.D., Ph.D.[f,*], ASMBS Artificial Intelligence and Digital Surgery Task Force

- 36 clinical questions derived from published guidelines
- Three LLMs prompted identically
- Blind, duplicate grading by bariatric surgeons

Model	Likert (1–5)	Appropriate
ChatGPT-4	4.46	**85.7%**
Bard (Gemini)	3.89	74.3%
Bing	3.11	25.7%

Lee Y, et al. Surg Obes Relat Dis 2024;20:603–608. $p < 0.001$ across models.

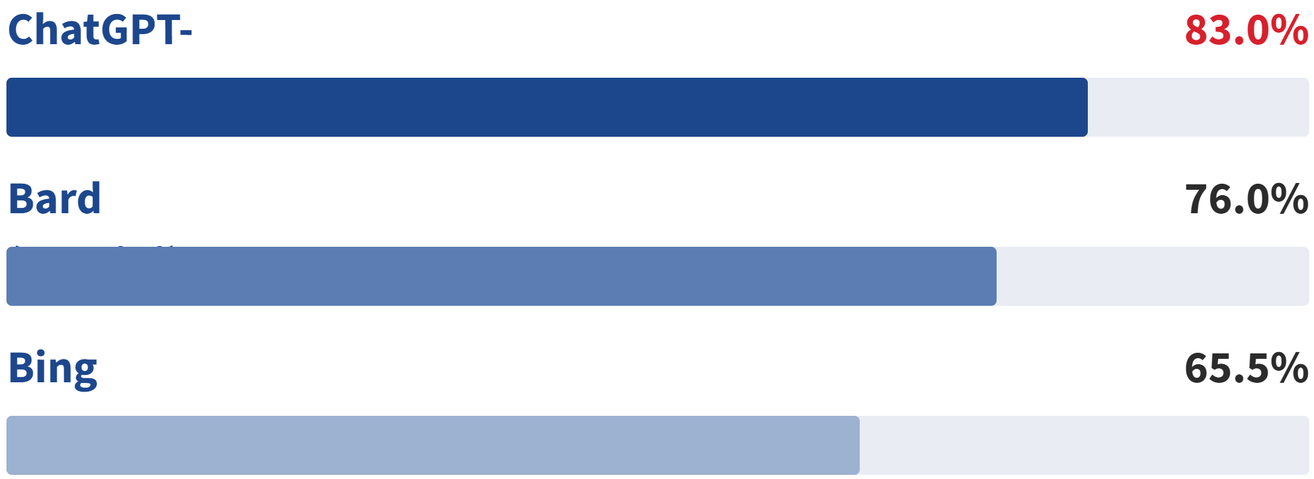
Can LLMs pass an MBS textbook exam?

ASMBS Guidelines/Statements

Performance of artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and Bard in the American Society for Metabolic and Bariatric Surgery textbook of bariatric surgery questions

Yung Lee, M.D., M.P.H.[a,b], Léa Tessier, M.D.[a], Karanbir Brar, M.D.[a], Sarah Malone, B.H.Sc.[a], David Jin, B.H.Sc.[a], Tyler McKechnie, M.D., M.Sc.[a], James J. Jung, M.D., Ph.D.[c], Matthew Kroh, M.D.[d], Jerry T. Dang, M.D., Ph.D.[d,*], ASMBS Artificial Intelligence and Digital Surgery Taskforce

200 multiple-choice questions from the ASMBS Textbook of Bariatric Surgery, 2nd edition. Correct answers, $p < 0.001$ across models.

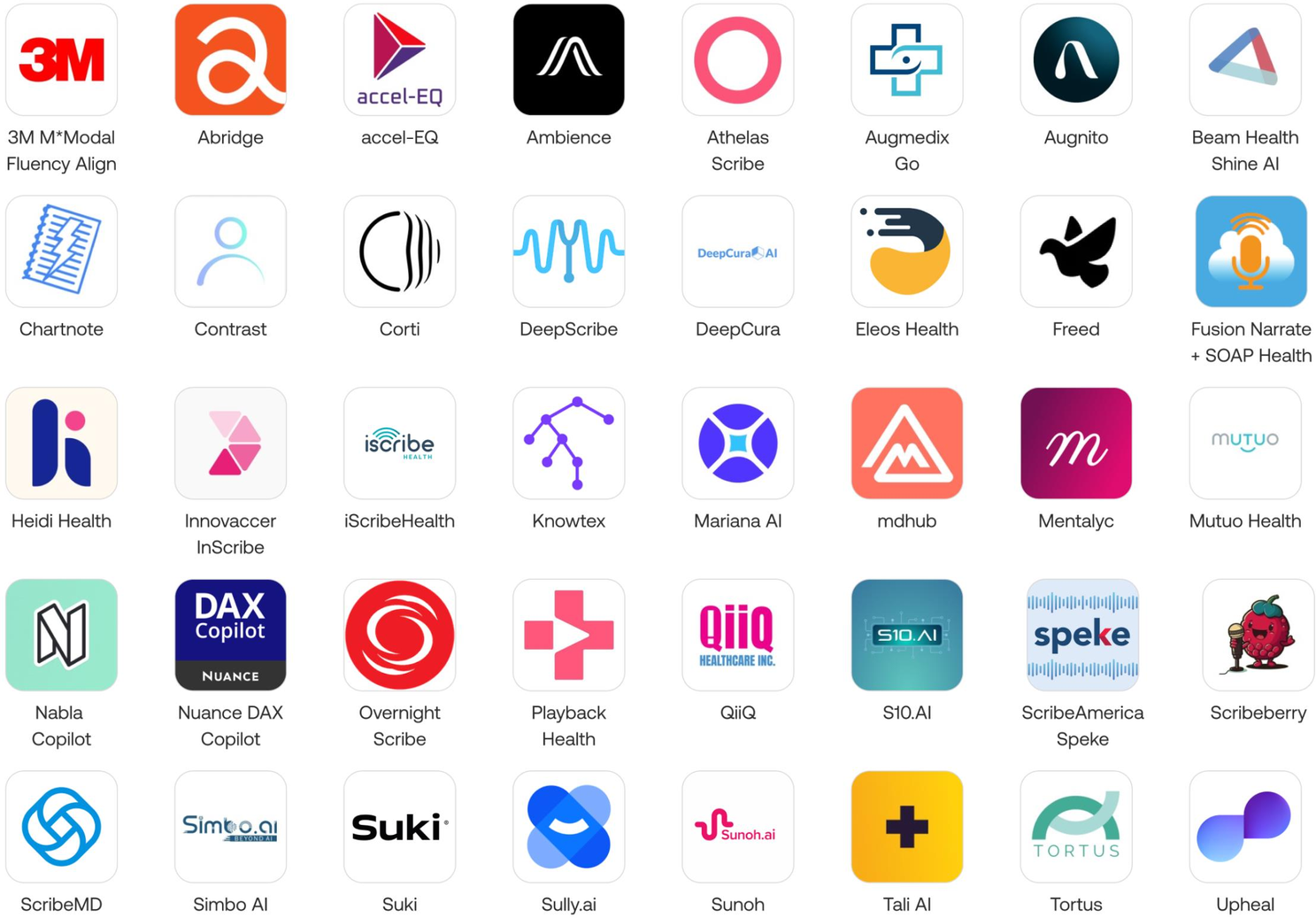


Lee Y, et al. Surg Obes Relat Dis 2024;20:609–613.

Ambient AI Scribes in the Bariatric Clinic

Elion

AI Scribe Category



Ambient AI Scribes: Observational Benefits

Less documentation time

−20.4%

time-in-notes per appointment (Duggan) · −0.57 min/note (Ma)

Less after-hours “pajama time”

−30.0%

per workday (Duggan) · −5.17 min/day (Ma)

Lower task load and burnout

−24.4 pts

task load; burnout −1.94 pts, both $p < .001$ (Shah)

Improved engagement

“More present with patients”

clinician-reported (Duggan)

Ma SP, et al. JAMIA 2025;32:381–385 · Duggan MJ, et al. JAMA Netw Open 2025;8:e2460637 · Shah SJ, et al. JAMIA 2025;32:375–380.

Ambient AI Scribes: First Randomized Trial

238 physicians **14** specialties · **16%** surgeons **~48,000** visits **3 arms**: DAX · Nabla · usual care **2 months**

Outcome vs usual care	**DAX Copilot** (n = 79)	**Nabla** (n = 79)
Time-in-note primary · % change, month 2 vs baseline	−1.7% (−9.4 to +5.9)	**−9.5%** (−17.2 to −1.8; P = 0.02)
Burnout composite (Mini-Z 2.0) scale 10–50 · higher is better	**+2.8** (+1.3 to +4.4)	**+2.7** (+1.1 to +4.2)
Physician task load scale 0–400 · lower is better	**−40** (−72 to −8)	−32 (−64 to +0.4)
Work exhaustion (PFI-WE) scale 0–4 · lower is better	**−0.32** (−0.55 to −0.08)	−0.23 (−0.46 to +0.01)
Clinically significant inaccuracies 1–5 Likert · 3 = “occasionally”	2.7 (2.4 to 3.0)	2.8 (2.6 to 3.0)

Red = 95% CI excludes zero · Scribe used in 33% (DAX) / 30% (Nabla) of visits · 1 grade-1 adverse event

Lukac PJ, et al. Ambient AI Scribes in Clinical Practice — A Randomized Trial. NEJM AI 2025;2(12). doi:10.1056/AIoa2501000

Ambient AI Scribes: Challenges

Variable accuracy

Manual review and editing remain essential; physician views on accuracy largely negative (Shah)

Note bloat

Note length increased 20.6% (Duggan)

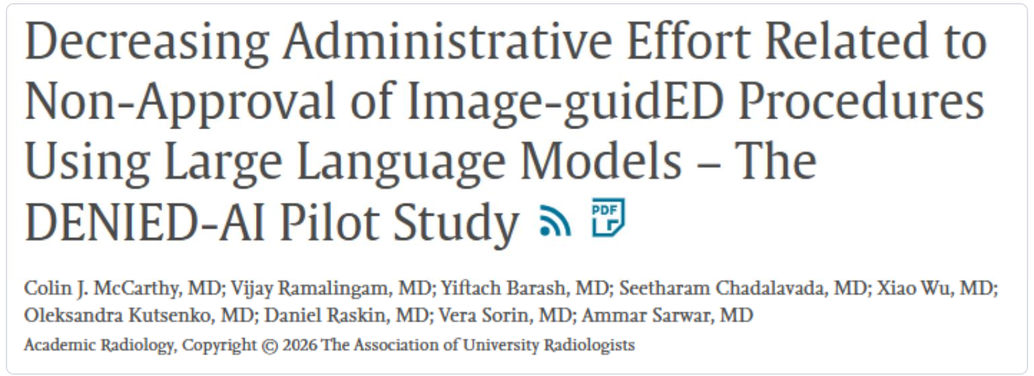
Contextual limits

Reduced accuracy in complex or non-English encounters

Heterogeneous benefit

Utility and time savings vary widely among users (Ma); tool usage in the RCT was only ~30% of visits

Can LLMs write insurance appeal letters?



Decreasing Administrative Effort Related to Non-Approval of Image-guidED Procedures Using Large Language Models – The DENIED-AI Pilot Study

Colin J. McCarthy, MD; Vijay Ramalingam, MD; Yiftach Barash, MD; Seetharam Chadalavada, MD; Xiao Wu, MD; Oleksandra Kutsenko, MD; Daniel Raskin, MD; Vera Sorin, MD; Ammar Sarwar, MD

4 LLMs (Claude 3.5, Nova Pro, Llama-3.1-70B, GPT-4o) · zero-shot, few-shot and RAG prompting · radiology procedure denials

73%
of letters useful as templates

71%
needed <10 min of editing

33%
of assessments flagged hallucinations

80%
of references fabricated by offline models (vs 29% GPT-4o, $p < .001$)

McCarthy CJ, et al. DENIED-AI Pilot Study. Academic Radiology 2026. Promising starting templates — human review required.

03

The Future of AI in MBS

International Expert Consensus on AI in MBS

Modified Delphi · 68 bariatric surgeons · 35 countries · 28 statements · ≥70% threshold · all reached consensus in 2 rounds

Surgical Education & Training

Objective skill assessment, personalized feedback, streamlined education

Clinical Decision-Making & Planning

Aids patient and procedure selection — but surgeon oversight of final decisions is essential

Prediction & Follow-Up

Predicting weight loss, complications and remission; long-term follow-up

Cost & System Support

EHR integration to identify social determinants, optimize processes and reduce cost

> Ethics are paramount: adherence to guidelines, inclusion in patient consent, and clear accountability for AI-driven decisions. Mandatory AI education in surgical curricula is coming.

Kermansaravi M, et al. Sci Rep 2025;15:9312.

Conclusions

THE PROMISE — EFFICIENCY & SUPPORT

Patient education

LLMs give accurate, comprehensive, reproducible answers to common bariatric questions

Clinician aid

Strong on guideline-based questions and textbook exams

Administrative efficiency

Ambient scribes cut documentation time, after-hours EHR work and burnout — now with randomized evidence

THE PERIL — INACCURACY & MISUSE

Safety-critical omissions

Preferred by patients, yet an LLM answer omitted post-op fluid guidance; scribes “occasionally” drop or misattribute content

Hallucination

A third of appeal letters flagged hallucinations; offline models fabricated 80% of references

Weak evidence base

58 studies, zero prospective; randomized trials only appearing in 2025 — surgeon oversight remains essential

IFSO

XXIX IFSO World Congress

1-4 September 2026
Toronto, Canada

Thank you

Questions?

Jerry Dang, MD, PhD · Cleveland Clinic
dangj3@ccf.org

XXX IFSO WORLD CONGRESS

"Advancing Multimodal Obesity Care"

24-27 August 2027

ifso2027.org

IFSO
PARIS 2027

See you in

PARIS