Open Book Test: Measuring OpenEvidence Performance on EdAI Certification Questions

Background: Recent research (Dec 2025) showed OpenEvidence achieves 100% accuracy on USMLE but only 31-41% on specialty board questions from the MedXpertQA dataset.

Methods:

- Generated 268 certification questions using EdAI from specific medical articles
- Removed 83 questions where OpenEvidence lacked access to source material
- Tested remaining 185 questions on OpenEvidence platform under "open book" conditions
- Documented accuracy across 7 medical specialties

Here's the specialty-by-specialty breakdown:

Specialty	***Correct***	***Total***	***Accuracy***
Obesity Medicine	19	24	79.17%
Cosmetic Surgery (ABCS)	24	32	75.00%
Pediatric Neurosurgery	30	40	75.00%
–Gynecologic Oncology	12	17	70.59%
Urologic Oncology	13	20	65.00%
Laser Surgery (ABLS)	18	32	56.25%
Facial Cosmetic Surgery (ABFCS)	11	20	55.00%
Overall	127	185	68.65%

Key Findings:

1. **Open book failure:** Despite direct access to source articles (83 questions excluded to ensure this), accuracy was only 68.65%
2. **Below passing threshold:** Medical students require 70%+ to pass certification

EdAI Systems. Inc | 677 Ala Moana Blvd. #1011 | Honolulu, HI 96813
john@edaisystems.com | edaisystems.com | 808-352-1999

3. **Specialty variation:** 24-point spread (55-79%) demonstrates inconsistent reasoning
4. **Severe failures:** Two specialties (ABLS, ABFCS) failed nearly half of questions despite source access

Implications: General-purpose medical AI cannot serve professional certification markets. Content access ≠ content control. The ability to retrieve information does not equal the ability to reason through certification-level complexity.

EdAI Differentiation: Content-driven intelligence architecture generates questions that require synthesis and clinical reasoning beyond RAG retrieval capabilities. This creates a defensible moat in high-stakes certification domains.