Examining Large Language Models on the Task of Improving Readability of Patient Education Materials for Hernia Surgery

Amber Chen-Goodspeed MD, Julie Hong, MD, Phillip Hwang MD, Aushja Syed MD, Victoria Ende MD, Steven Chao MD

BACKGROUND

- Patient education materials (PEMs) are essential for informed consent and shared decision-making.
- The American Medical Association recommends materials be written at a 6th-grade reading level.
- Studies show most surgical PEMs exceed this level.
- Large Language Models (LLMs) may simplify medical content.
- However, concerns remain regarding:
  - Replicability
  - Reliability
  - Risk of misinformation or hallucination

Study Aim

To evaluate whether three LLM platforms can reliably and repeatedly simplify a ventral hernia PEM to a sixth-grade level using a standardized prompt.

METHODS

Design: Comparative reproducibility study of three LLM platforms.
Data Source: SAGES Ventral Hernia PEM
Baseline FKGL: 8.0
Procedure

- PEM Rewritten 5× per platform using identical prompt
- Platforms:
  - ChatGPT-5
  - NotebookLM
  - OpenEvidence

Outcomes:

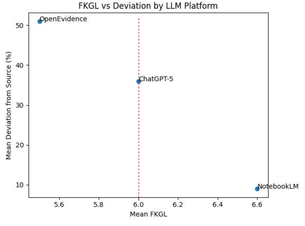
- FKGL
- Processing time
- Deviation from source (%)
- Qualitative error patterns

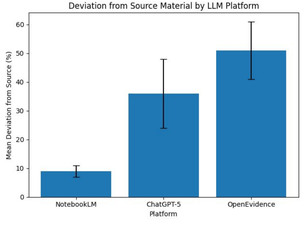
Sample output:

Original (SAGES); FKGL 8.0	NotebookLM Version; FKGL 6.0
"A hernia is a **gap, space or hole** in the **strong tissue (fascia)** that holds muscles in place. Hernias occur if the layers of fascia in your **abdomen** become weak. Then the inner lining of the **abdomen** pushes through the gap. It forms a **small sac**, like a balloon."	"A hernia is a **hole** in the strong **muscle** of your **belly.** Parts of your body, like your intestines, can push through this hole. This can create a **bulge** under your skin."

RESULTS

- **All platforms improved readability from original content FKGL 8**
- Closed-source (NotebookLM) vs open-source (ChatGPT and OpenEvidence) design influenced variability.





Qualitative Behavior Patterns

OpenEvidence	NotebookLM	ChatGPT-5
• Introduced outside references despite prompt restriction • Lowest FKGL	• Fastest • Most replicable • Oversimplified terminology ("muscle" for "fascia")	• No citations • Added accurate but unreferenced information

DISCUSSION/CONCLUSION

- LLMs can improve PEM readability. All LLMs successfully reduced readability to ~6th-grade level.

<u>However:</u>

- Readability ≠ accuracy.
- Replicability varies significantly by model.

NotebookLM demonstrated:

- Fastest processing
- Lowest deviation
- Most reproducible outputs

- Replicability appears strongest in single-source models (NotebookLM).

References

1. SAGES. "Laparoscopic Ventral Hernia Repair Information from SAGES." *Society of American Gastrointestinal and Endoscopic Surgeons (SAGES)*, 19 Feb. 2022, www.sages.org/publications/patient-information/patient-information-for-laparoscopic-ventral-hernia-repair-from-sages/. Accessed 15 Sept. 2025.

For questions please contact: Amber Chen-Goodspeed, MD, PGY-4; amber.chengoodspeed@gmail,