Cloud vs On-Premises LLMs for Urgent Patient Message Screening

Valdery Moura Junior, PhD, MBA[1,2]; Susanna Gallani, PhD[2]; Lara Basovic, MD[1]; Majed Alomar, MBBS[1]; Jason You, MD, PhD[1]; Elyse Park, PhD[1]; Gaurdia Banister, PhD, RN[1]; Louisa Sylvia, PhD[1]; Shawn Murphy, MD, PhD[1]; Peter Hadar, MD[1]; Lidia Moura, MD, PhD, MPH[1]

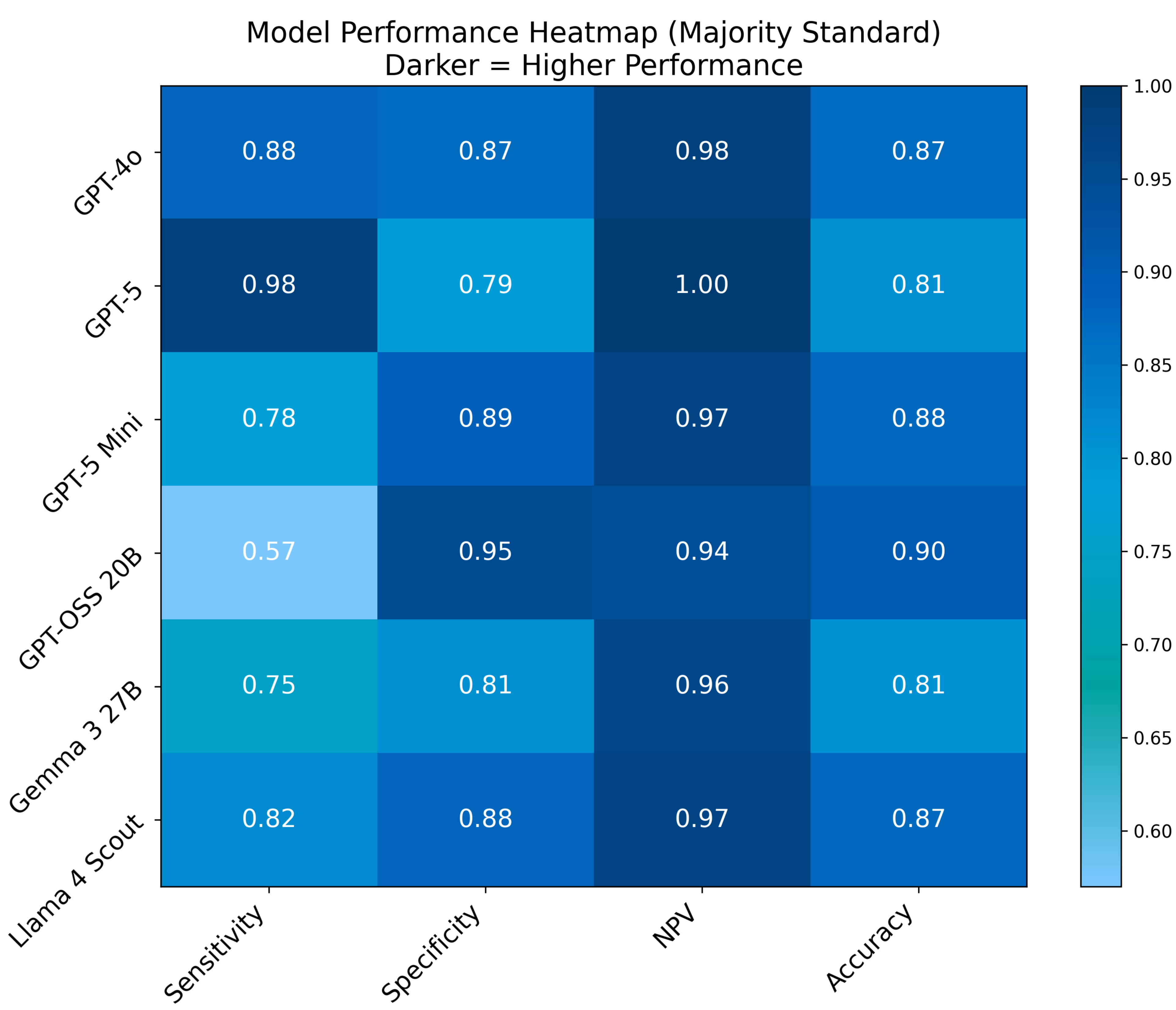
[1] Massachusetts General Hospital, Harvard Medical School, Boston, MA; [2] Harvard Business School, Boston, MA

Massachusetts General Hospital
Founding Member, Mass General Brigham

HARVARD MEDICAL SCHOOL

Introduction

Patient portal messaging has become a core channel for outpatient neurology care, particularly in high-risk specialties like epilepsy (1,2). Epilepsy affects over 3 million Americans and often requires complex, time-sensitive management of seizures, life-threatening medication side effects, and urgent prescription needs. The exponential growth in message volume, combined with a national shortage of neurologists, has placed increasing demands on both subspecialty and primary care providers. Urgent messages are frequently buried among routine inquiries, delaying care and contributing to patient dissatisfaction and clinician burnout (1,3). We aimed to evaluate the feasibility of using large language models (LLMs) for automated message triage as a scalable solution in outpatient care, including comparative performance of cloud vs on-premises models and implications for PHI handling.

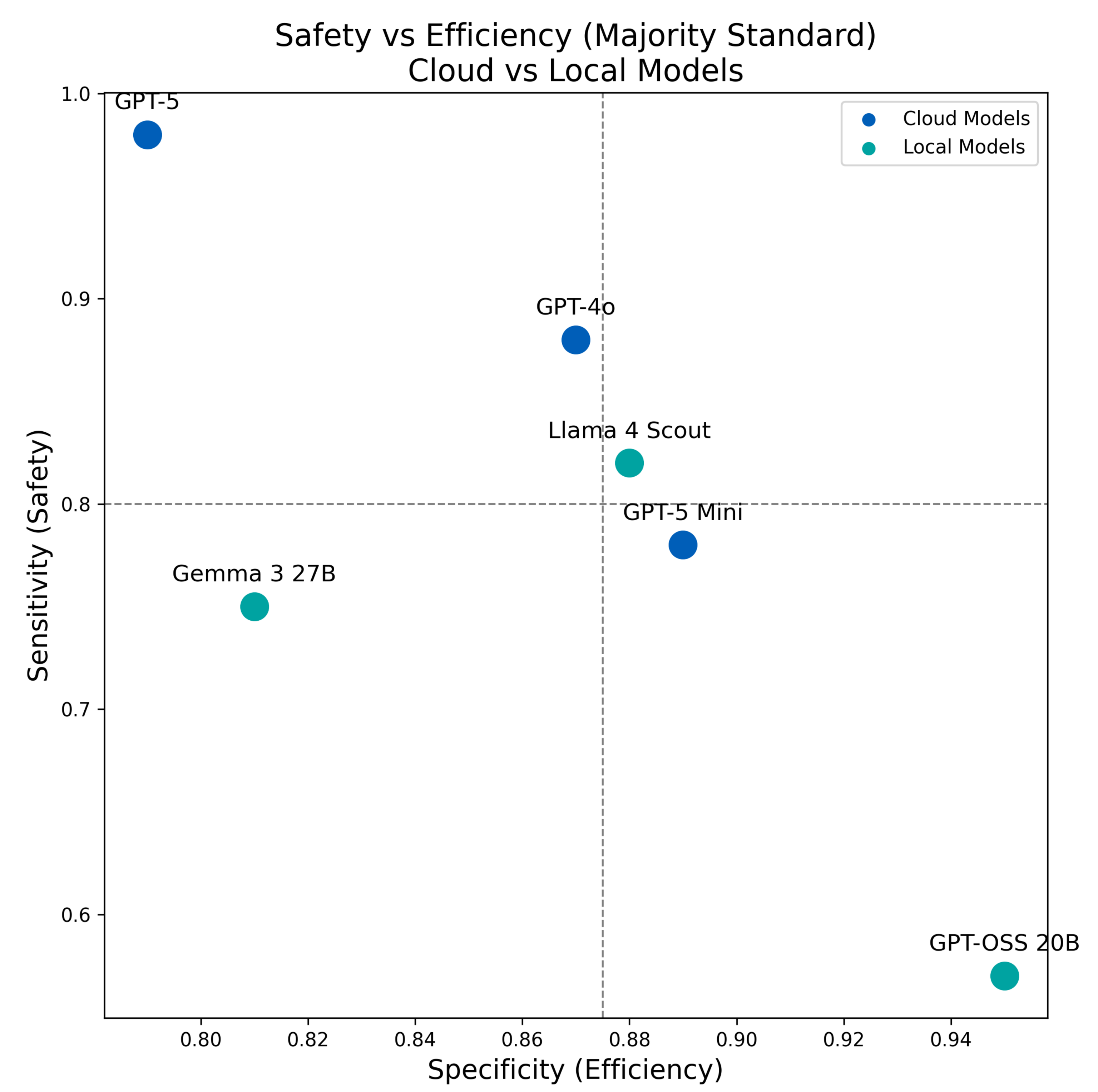


Methods

We conducted a retrospective diagnostic accuracy study to evaluate the performance of large language models (LLMs) in classifying the urgency of outpatient neurology patient messages. During a structured training phase (n=50), five fellowship-trained epilepsy neurologists achieved high inter-rater reliability (κ ≥ 0.80) on urgency labels. They then annotated 503 de-identified patient portal messages from a tertiary epilepsy clinic using a standard operating procedure (SOP). We benchmarked six LLMs - three cloud-based (GPT-4o, GPT-5, GPT-5 mini) and three on-premises (Llama 4 Scout, GPT-OSS 20B, Gemma 3 27B) - in a zero-shot configuration with deterministic decoding where supported. Recognizing that safe clinical deployment requires near-perfect sensitivity and negative predictive value (NPV), along with sufficient specificity to avoid alert fatigue, we evaluated trade-offs across safety, efficiency, and data governance. Primary outcomes included sensitivity and NPV under both Unanimous [5/5 rater agreement] and Majority physician consensus standards [≥3/5 rater agreement].

Results

Under the strict Unanimous consensus standard, all models except GPT-OSS 20B achieved perfect sensitivity (1.00) and NPV (1.00), missing no urgent messages. GPT-OSS 20B was the exception, with lower sensitivity (0.69 [95% CI: 0.39–0.91]) but high specificity (0.98 [0.96–0.99]) and accuracy (96.9% [94.6–98.4]). Under the more pragmatic Majority standard, GPT-5 prioritized safety, achieving the highest sensitivity (0.98 [0.91–1.00]) and NPV (1.00 [0.98–1.00]). GPT-4o and Llama 4 Scout offered more balanced profiles, combining strong specificity (0.87 [0.83–0.90] and 0.88 [0.85–0.91]), high NPV (0.98 and 0.97), and solid accuracy (86.9% [83.6–89.7] and 87.3% [84.0–90.1]). GPT-OSS 20B delivered the highest specificity (0.95 [0.92–0.97]) and overall accuracy (90.3% [87.3–92.7]) but showed lower sensitivity (0.57 [0.43–0.69]). Gemma 3 27B demonstrated intermediate performance, with sensitivity of 0.75 [0.62–0.85], specificity of 0.81 [0.78–0.85], and accuracy of 80.7% [77.0–84.1]. Overall, local models—including Llama 4 Scout and GPT-OSS 20B - closely approximated the performance of their cloud-based counterparts, with sensitivity trade-offs for GPT-OSS 20B.



Discussion and Conclusions

These findings demonstrate that both cloud and local LLMs are viable for urgent message screening in outpatient neurology. GPT-5 prioritized safety (sensitivity ~0.98, NPV ~1.00); GPT-4o optimized efficiency with balanced sensitivity/specificity. Local models - Llama 4 Scout, GPT-OSS 20B, and Gemma 3 27B - approached cloud-level performance, offering scalable, privacy-preserving deployment paths for PHI-sensitive workflows.

While GPT-5 is well suited for safety-critical settings, GPT-4o and Llama 4 Scout balanced specificity and alert burden without compromising rule-out reliability. GPT-OSS 20B delivered the highest accuracy and lowest false-positive rate but requires human oversight due to lower sensitivity. Next steps include prospective validation, testing across other hospital services, and reproducibility checks under deterministic inference where supported.

References: 1. Nath B, Williams B, Jeffery MM, O'Connell R, Goldstein R, Sinsky CA, et al. *Trends in electronic health record inbox messaging during the COVID-19 pandemic in an ambulatory practice network in New England*. JAMA Network Open. 2021; 2. Holmgren AJ, Thombley R, Sinsky CA, Adler-Milstein J. *Changes in physician electronic health record use with the expansion of telemedicine*. JAMA Intern Med. 2023; 3. Baxter SL, Saseendrakumar BR, Cheung M, Savides TJ, Longhurst CA, Sinsky CA, et al. *Association of electronic health record inbasket message characteristics with physician burnout*. JAMA Network Open. 2022.