Evaluating AI Medical Scribes: Performance, Precision, and Promise

Assessing Om Medical's AI Scribe Against Human and Commercial Standards (OM-003)

Study conducted at Stony Brook University Hospital, Clinical Sim Center, Oct 10, 2024
Study completed: Jan 21, 2025

Acknowledgement

This study was conducted at the clinical simulation center of Stony Brook University Hospital, Renaissance School of Medicine, with the help of volunteers, case participants, and contributors including: Matthew Tharakan, M.D., Lyncean Ung, M.D., and Rachel Wong, M.D.

*Acknowledgement does not imply affiliation or endorsement.

Introduction

Clinical documentation is a critical yet time-consuming component of healthcare delivery, often contributing to physician burnout and decreased patient interaction time (Shanafelt et al, 2017; Sinsky et al, 2016). Ambient AI scribe systems have recently emerged as promising tools that automatically transcribe and structure clinical encounters, thereby alleviating the administrative burden on clinicians and potentially improving both documentation quality and workflow efficiency (Davenport & Kalakota, 2019). These systems leverage advances in artificial intelligence—particularly large language models (LLMs)—to capture patient encounters in real time, generate coherent clinical notes, and organize complex information with minimal human intervention.

Despite the promising efficiency gains, concerns regarding the accuracy, reliability, and potential for information "hallucination" remain. Early evaluations indicate that while AI-generated notes can match or exceed the organizational quality of human scribes in many instances, they occasionally introduce errors or misinterpret clinical nuances (Challen et al, 2019). Moreover, the rapid evolution of reasoning models—such as GPT-o1 and its contemporaries—raises important questions about which models are best suited for this task. In this study, we compare Om Medical's ambient AI scribe with several commercial solutions and an experienced human scribe using a range of simulated clinical scenarios.

Methodology

This study was conducted at the Stony Brook University Renaissance School of Medicine (RSOM) Sim Center with volunteer participants enacting six clinical scenarios: a simple primary care case, complex primary care case, psychiatric encounter, post-operative follow-up, trauma case, and inpatient encounter. Each scenario involved scripted dialogues that were recorded simultaneously by devices running AI scribe applications. Specifically, we tested systems from **Nuance DAX**, **Nabla**, **Abridge**, **Suki**, **Ambience**, and **Om Medical**. Also included was an experienced human scribe with over 5 years of scribing experience in a professional clinical setting. Case dialogue was designed to reflect the variation and unpredictability of real-world clinical encounters.

Om medical

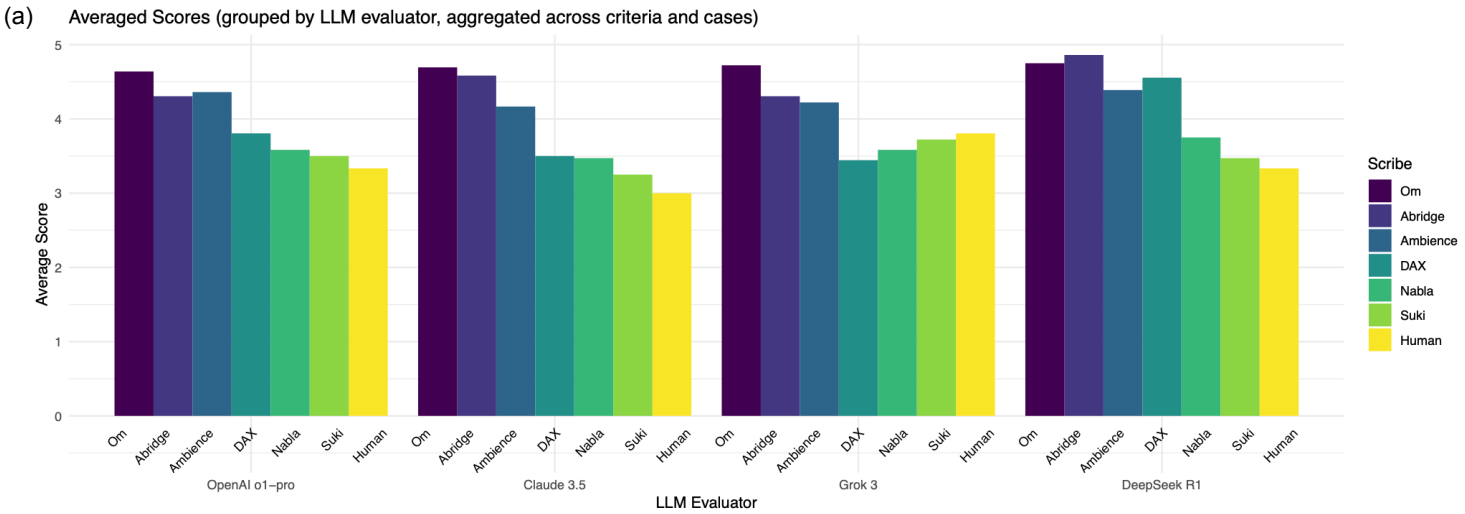
To evaluate the generated clinical notes, we developed a rubric that assesses six dimensions: completeness and relevance of clinical content; organization and clarity; accuracy and specificity; handling of complexity and interruptions; conciseness and readability; and adaptability to varied clinical workflows. The rubric, which is appended to this paper, specifies subcriteria for each dimension, ensuring that both objective data (e.g., correct medication dosages and lab values) and subjective aspects (e.g., logical flow and clarity) are evaluated.

Rather than relying on human evaluators for scoring, we employed LLMs to score clinical notes —namely OpenAI o1-pro, Grok 3, Claude 3.5, and DeepSeek R1. These models, which represent the latest reasoning capabilities at the time of writing, were chosen for their ability to reduce evaluator bias and provide standardized, reproducible assessments. For each simulated case, each LLM evaluator was prompted with the case dialogue, the evaluation rubric, and the corresponding scribe-generated notes from each participant. The LLM was prompted to return a structured object representing scores across all rubric criteria.

For each evaluation, only the first result returned by an LLM was used in our analysis. Re-running of the LLM was avoided except under extenuating circumstances. Specifically, in one instance, an LLM misinterpreted the prompt and erroneously split the output for the human scribe into two separate results (labeled human_1 and human_2), as well as the output for Nabla (labeled nabla_1 and nabla_2). This anomaly was resolved by re-running the prompt, which then produced the expected output format. Aside from this instance, all other 23 LLM evaluations were executed once with no repetition. The resulting JSON objects were aggregated and subsequently processed in R for statistical analysis and visualization.

Results

Om achieved consistently high scores across OpenAI o1-pro, Claude 3.5, and Grok 3 (4.64, 4.69, and 4.72, respectively), and ranked second (4.75) behind Abridge (4.86) under DeepSeek R1's evaluation. Om was thus ranked first by three of the four LLM evaluators (Figure 1).



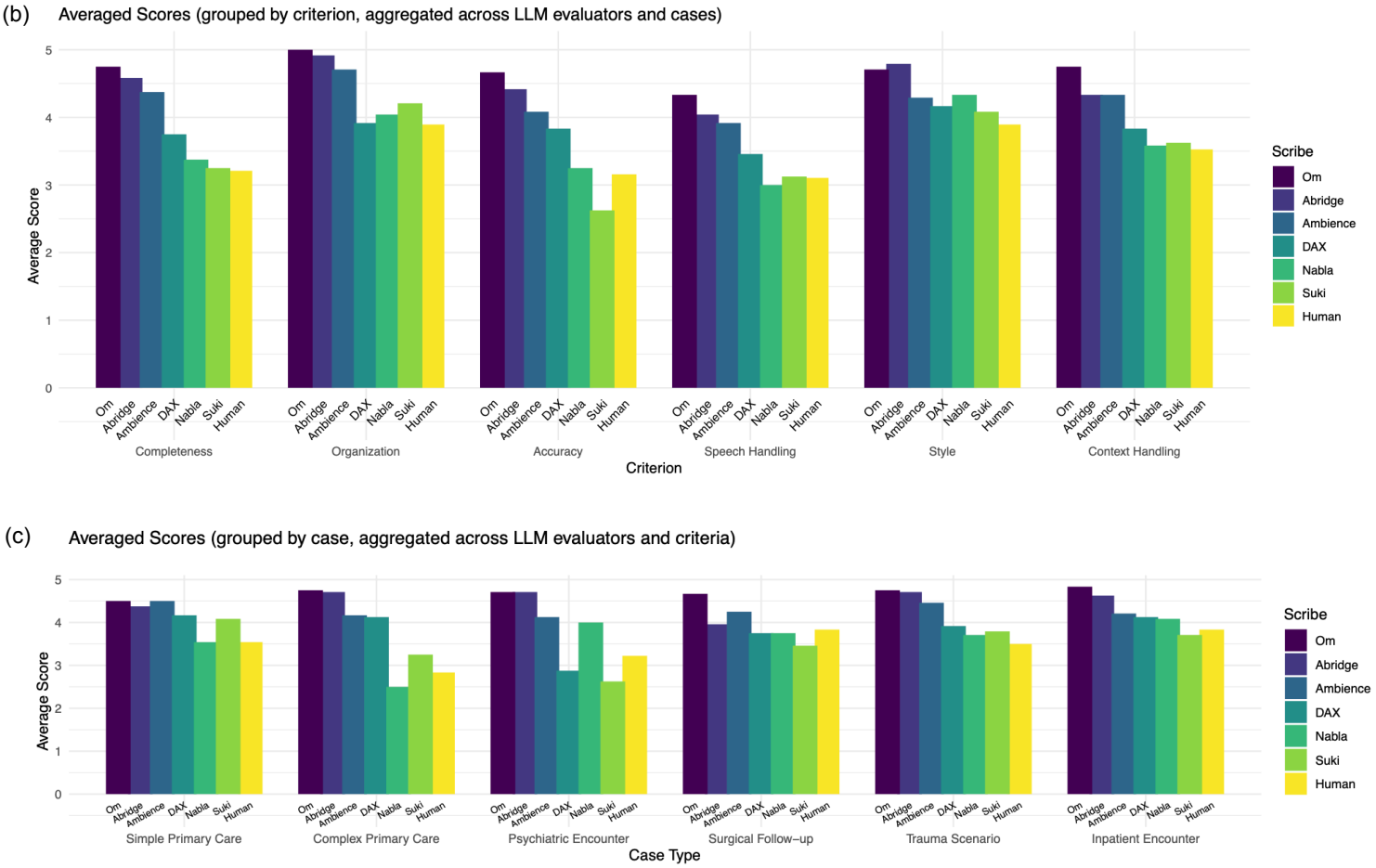


Figure 1. Scribe Performance by LLM Evaluator, Criterion, and Case. (a) Average scores grouped by large language model (LLM) evaluators. (b) Scores aggregated across rubric criteria. (c) Scores aggregated across clinical scenarios.

Analysis per rubric criterion revealed that Om led in Completeness (4.75), Accuracy (4.67), Organization (5.00), Context Handling (4.75), and Speech Handling (4.33). Abridge exceeded Om in Style (4.79 vs. 4.71). Om maintained the highest composite scores across all other metrics. Case-specific analyses show that Om held the highest or jointly highest score in all six scenarios, ranging from a mean of 4.50 in simple primary care to 4.83 in the inpatient encounter (Figure 2).

LLM evaluators that had "reasoning" ability (DeepSeek R1 and OpenAI o1-pro) exposed their chain of thought and could therefore be introspected for justifications of the outputed scores. Notable details mentioned include inaccurate transcription of medical terms (e.g, Nabla replaced "IV" with "MIV" in its note for case 1), misstatement of patient age by Suki (e.g, reports patient's age as 73 rather than 20 in case 3), and inability to deal with inconsistency (e.g, EMS reported left arm IV placement whereas nurse reported right arm placement in case 5; only Om refrained from specifying a laterality). Thus LLM-based scoring not only served as an unbiased evaluation of AI scribes, but also offered specific insight on scribe performance per case and per rubric criteria.

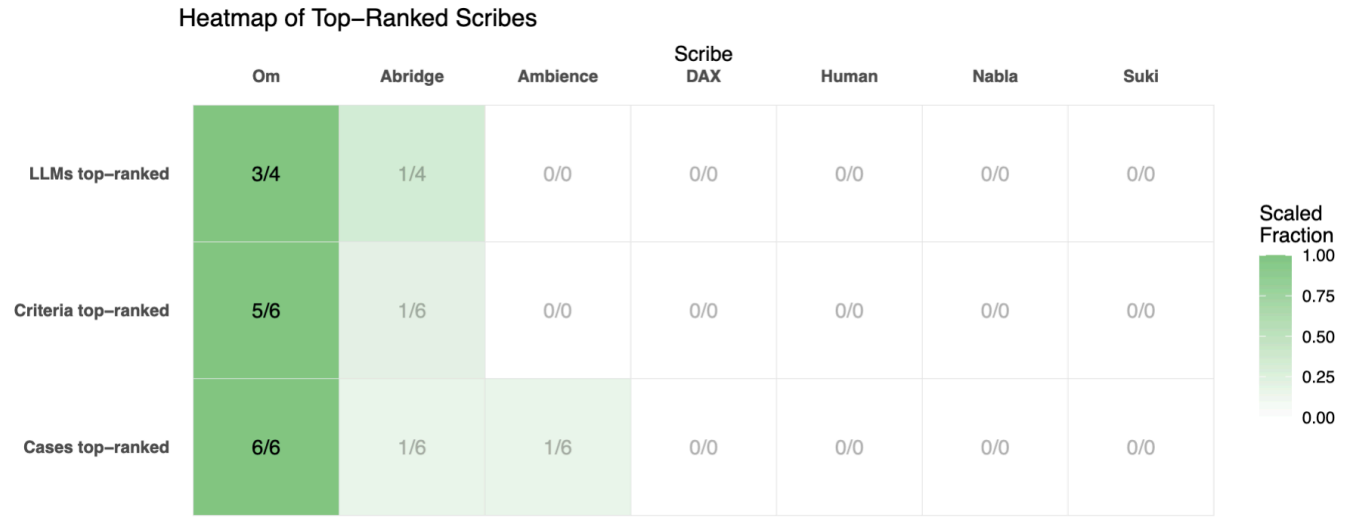


Figure 2. Frequency of Top-Ranking Across Scribe Evaluations. Heatmap summarizing how often each scribe achieved the top ranking across evaluations by LLM evaluator evaluation criterion, and clinical case. Darker shading indicates a higher proportion of top ranks.

Discussion

The present study evaluated seven scribe solutions—six AI-powered and one experienced human scribe—across multiple simulated clinical scenarios using automated, large language model (LLM)-based scoring. The results suggest that Om and Abridge maintained relatively higher performance than the other AI scribes (Ambience, DAX, Nabla, Suki) and the human scribe in most domains and scenarios. In particular, Om achieved the top or near-top aggregate scores for completeness, accuracy, organization, context handling, and speech handling, while Abridge outperformed Om in the style criterion. Cases involving complex or nuanced clinical interactions (e.g., trauma, psychiatric) tended to showcase the importance of robust speech handling, context retention, and correct capture of clinical details, dimensions in which Om and Abridge performed well.

Although AI scribes have demonstrated promising performance in these simulated environments, the findings are derived from a controlled setting without direct physician input or real-world edge cases. The decision to rely on LLM-based evaluators reduces certain forms of inter-rater bias inherent to human judging, but also introduces potential biases stemming from the prompt designs and internal weighting of each LLM. Still, the broad agreement among multiple LLM evaluators in their scribe rankings speaks to the method's validity and reproducibility.

Several important limitations should be noted. First, no direct physician evaluation was included to confirm whether the documented notes sufficiently meet clinical standards or improve actual workflow. Second, time metrics were not measured; thus, it remains unclear whether certain scribe solutions offer more efficient note generation in practice. Third, only simulated cases were used, which may not fully capture the complexities and unpredictable nature of real-world patient encounters. Fourth, evaluation is based on subjective reasoning by models and its scoring may not reflect the preferences of a human reader. Future studies may incorporate real clinical workflows, external physician panels, and robust time and cost analyses to more comprehensively assess the efficacy of these AI scribes.