LangGraph-Orchestrated Agentic AI Scribes for Home Health Documentation: System Architecture and Fact-Level Evaluation

Pallavi Gupta, PhD, MS[1], Zhihong Zhang, PhD, RN[1,2], William Ho, BS, MS[3]; Yu-Wen Chen, BS, MS[4]; Sasha Vergez, MPH[5]; Margaret McDonald, MSW[5]; Zoran Kostic, PhD[3]; Julia Hirschberg, PhD[4]; Maxim Topaz, PhD, MA, RN[1,2,5]

[1]School of Nursing, Columbia University, New York, NY; [2]Data Science Institute, Columbia University, New York, NY;[3]Department of Electrical Engineering, Columbia University, New York, NY; [4]Department of Computer Science, Columbia University, New York, NY; [5]Center for Home Care Policy & Research, VNS Health, New York, NY.

Background: Patient–clinician conversations during home health care (HHC) visits are information-dense, yet important problems and interventions are not fully captured in documentation. One study found, ~50% of patient problems and ~20% of interventions discussed during HHC visits never reached the EHR[1], undermining care continuity, quality measurement, and analytics, while adding documentation burden for clinicians managing multiple visits in a day. Another study found, adding conversational data to risk models improved prediction of emergency visits and hospitalizations by 26%[2]. LLM-based AI scribes are already being used to mitigate above challenges, but single-pass generation can blur evidence vs inference, increasing errors, especially in longer conversations during HHC[3]. In addition, common evaluations using BLEU/ROUGE or human ratings reward lexical overlap and surface fluency but fail to detect in-depth missing or fabricated clinical facts[4]. We address these gaps with a verifier-augmented, multi-step LangGraph-orchestrated agentic AI scribe and fact-level evaluation.

Methods: Fifteen HHC patient-clinician audio recordings (>8 hours, recorded by VNS health-one of the largest HHC provider in the US), were stratified by duration into short (10–25 min; n=5), medium (25–45 min; n=5), and long (45–70 min; n=5) encounters. Audio recordings were transcribed with OpenAI Whisper (medium) and diarized using PyAnnote Audio; speaker turns were labeled Patient/Clinician via GPT-4o role classification, yielding labeled transcripts. Transcripts were fed into a LangGraph-orchestrated agentic pipeline to generate SOAP (Subjective, Objective, Assessment, Plan) notes. The agentic pipeline comprised of four parallel section-specific drafting nodes (S: patient-reported symptoms; O: clinician observations; A: clinician assessment; P: plan of action) with anti-fabrication guardrails and clinical taxonomy constraints, followed by a merge node that combined sections into a SOAP draft and a verifier node that reviewed the draft against the transcript to identify and remove transcript-unsupported statements. The pipeline was run with two reasoning-capable model configurations, GPT-5 and GPT-5.1 (temperature=0, reasoning effort=medium), three times for each encounter-model configuration pair (yielding total 15*2*3= 90 notes). For the gold standard, two clinicians extracted a total 428 consensus atomic clinical facts from 15 transcripts (e.g., dizziness ×3 days; BP 142/88; follow-up scheduled), decomposing compound statements into separate facts. We manually compared AI generated SOAP notes against gold standard, and evaluated using metrics:precision, recall, hallucination rate (100-precision), omission rate (100-recall), run-to-run consistency (recall standard deviation across three runs to quantify consistency in clinical facts captured), and conciseness (output tokens/note).

Results: Across 15 encounters, the two model configurations exhibited a precision-recall tradeoff: GPT-5 configuration achieved higher precision (94.48%) with a correspondingly lower hallucination rate (5.52%) and lower recall (95.3%) with higher omission rate (4.7%), whereas GPT-5.1 configuration achieved higher recall (97.36%) with fewer omissions (2.70%) but lower precision (88.91%) with a higher hallucination rate (11.09%). Encounter duration amplified this tradeoff. In short encounters, both model configurations exceeded 95% precision and 94% recall. In medium encounters, GPT-5.1 configuration displayed higher recall than GPT-5 configuration (97.24% vs. 93.34%) with modestly lower precision (91.89% vs. 94.15%). In long encounters, GPT-5 configuration maintained 93.91% precision (hallucination rate 6.09%) while GPT-5.1 configuration precision declined to 79.63% (hallucination rate 20.37%), despite near-ceiling recall for both model configurations (GPT-5: 98.30%; GPT-5.1: 99.35%). Run-to-run variability remained low across all strata (≤ 0.5%). Operationally, GPT-5.1 configuration generated 2.2× longer SOAP notes than GPT-5 configuration (by output tokens). Hallucinations were primarily clinically plausible inferences rather than clearly wrong, fabricated facts absent in transcript.

Conclusion: LangGraph-orchestrated, verifier-augmented LLM scribes produced SOAP notes with high recall and low run-to-run variability across both configurations. However, precision declined with encounter length, with hallucinations becoming increasingly common (reaching 20% for GPT-5.1 configuration) in long encounters. Hence, safe deployment of AI scribes for medically complex older adults in HHC requires fact-level evaluation, length-aware configuration/guardrails, transcript-grounded verification, and clinician-in-the-loop review before notes enter the EHR.

References:

1.Song J, Zolnoori M, Scharp D, Vergez S, McDonald MV, Sridharan S, Kostic Z, Topaz M. Do nurses document all discussions of patient problems and nursing interventions in the electronic health record? A pilot study in home healthcare. JAMIA open. 2022 Jul 1;5(2):ooac034.; **2**.Zolnoori M, Sridharan S, Zolnour A, Vergez S, McDonald MV, Kostic Z, Bowles KH, Topaz M. Utilizing patient-nurse verbal communication in building risk identification models: the missing critical data stream in home healthcare. Journal of the American Medical Informatics Association. 2024 Feb 1;31(2):435-44.; **3.**Krishna K, Khosla S, Bigham JP, Lipton ZC. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. InProceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) 2021 Aug (pp. 4958-4972); **4**.Ben Abacha A, Yim WW, Michalopoulos G, Lin T. An investigation of evaluation metrics for automated medical note generation. arXiv e-prints. 2023 May:arXiv-2305.