Identifying Goals-of-Care Discussions with Large Language Models

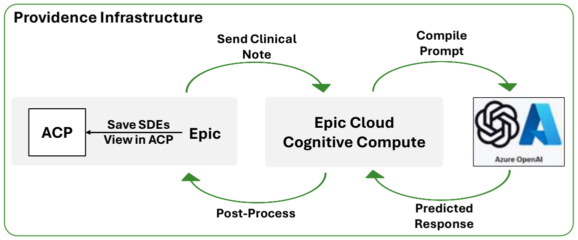
Providence
Institute for Human Caring

Matthew Gonzales, MD, FAAHPM; Deborah Unger, MD, FAAHPM; Guilford Parsons, MD, MS; Melissa Forbin, MBA; Suzanne Engelder, LCSW
Maulin Shah, MD; Canada Parrish, PhD; Azadeh Mobasher, PhD

Background and Context

- Patients experiencing serious illness and their families benefit from conversations about their **care values** and **goals**, which are essential to high-quality, equitable healthcare.
- At **Providence**, we've implemented system-wide efforts to improve documentation of these **Goals-of-Care** (**GoC**) discussions but tracking narrative, free-text entries in clinical notes remains challenging.
- Structured documentation tools exist but are often seen by clinicians as too rigid. **Traditional methods** such as manual review or rule-based systems struggle with cost, scalability, and accuracy.
- To address these limitations, we developed an **AI-powered** solution using general-purpose **large language models** (LLMs) to identify GoC conversations embedded in unstructured documentation.

Figure 1: GoC identification high-level architecture in Epic


Providence Infrastructure

ACP

Save SDEs
View in ACP

Epic

Send Clinical Note

Epic Cloud Cognitive Compute

Compile Prompt

Azure OpenAI

Predicted Response

Post-Process

Providence

120k CAREGIVERS
36k NURSES
25k PHYSICIANS
52 HOSPITALS
950 CLINICS
28.1m TOTAL PATIENT VISITS
1700+ PUBLISHED RESEARCH STUDIES
1 HEALTH PLAN
17 SUPPORTIVE HOUSING FACILITIES
1.2m HOME HEALTH VISITS
5m UNIQUE PATIENTS SERVED
HIGH SCHOOL NURSING SCHOOLS & UNIVERSITY
1.9m COVERED LIVES
$1.9b COMMUNITY BENEFIT

Email Authors

Development and Validation

- We used Azure OpenAI GPT models to detect the presence of key elements in GoC conversations after multiple rounds of prompt engineering, leveraging their strengths in contextual understanding and language processing.
- Annotation guidelines were established through multidisciplinary expert consensus, using **Labelbox** platform, resulting in substantial inter-rater reliability with **mean pairwise agreement** of **0.77**.
- We evaluated multiple proprietary LLMs, with **GPT-4o** achieving the highest performance on a dataset of **488** human-annotated clinical notes (Table 1).
- High **specificity** was prioritized for clinically relevant and actionable results. **Error analysis** was used to break the tie.

Table 1: Comparative performance of language models on GOC identification

Model	Specificity	Precision	F1 Score	Accuracy
GPT-4o-with-chunking	0.95	0.81	0.73	0.90
GPT-4o-without-chunking	**0.95**	**0.78**	**0.76**	**0.91**
GPT-4o-mini-with-chunking	0.93	0.76	0.76	0.91
GPT-4o-mini-without-chunking	0.88	0.62	0.71	0.87
GPT-4	0.92	0.72	0.76	0.90
GPT-3.5-Turbo	0.88	0.62	0.71	0.87

- Fairness and model performance was assessed across race and sex subgroups and **no statistically significant bias** was detected in model performance (Table 2).
- The final model was integrated into **Epic's Nebula** platform for real-time inference within existing predictive modeling infrastructure.

Table 2: Subgroup-specific performance for 488 human-annotated data

Subgroup		# Note	TP Rate	TN Rate	FP Rate	FN Rate	Specificity
Age in Years	0 – 44	66	1	1	0	0	1
	45 – 60	75	1	0.91	0.09	0	0.91
	61 – 70	97	0.64	0.96	0.04	0.36	0.96
	≥ 71	250	0.74	0.97	0.03	0.26	0.97
Sex	Male	245	0.72	0.96	0.04	0.28	0.96
	Female	243	0.78	0.96	0.04	0.22	0.96
Race / Ethnicity	White	327	0.73	0.96	0.04	0.27	0.96
	Black	27	1	1	0	0	1
	Other	134	0.82	0.96	0.04	0.18	0.96

Outcomes and Next Steps

- Following rigorous pre-production testing and institutional reviews across informatics, clinical leadership, and cybersecurity, the model was deployed in a pilot implementation across **four hospitals** in **December 2024**.
- **Epic's SmartData elements** were leveraged to monitor model usage and track documentation identified as containing GoC discussion.
- Identified notes are surfaced directly to clinicians via the **Advance Care Planning** (**ACP**) **Summary Report**, which serves as the centralized clinical reference for ACP and GoC documentation.
- Domain experts conduct **continuous evaluation** of stratified random samples of model predictions.
- Our data scientists employ **meta-prompting** strategies **LLM-as-a-judge** methodologies to support **expert adjudication**, inform **prompt engineering**, enabling systematic **identification of error patterns**.
- Using traditional methods, an average of **1,908** Goals-of-Care discussions per week are captured across **52** acute care hospitals. Within **the four pilot sites**, the LLM identified an average additional **14** documented conversations per week.
- We have governance approval to launch the model across the entire Providence health system.

Figure 2: AI-powered GoC identification pipeline is integrated into Epic ACP, complementing existing clinical workflow. This image is a property of Epic Systems Corporation.

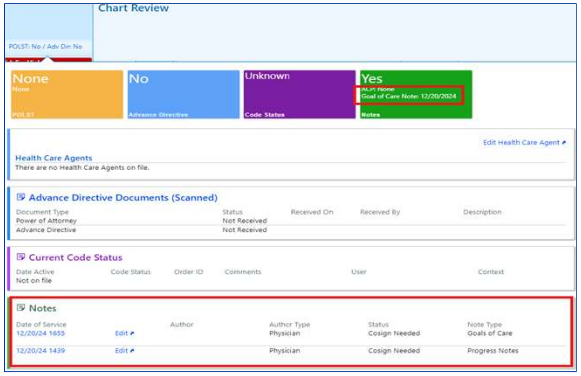


Chart Review

POLST: No / Adv Dir: No

None
None
POLST

No
Advance Directive

Unknown
Code Status

Yes
ACP: None
Goal of Care Note: 12/20/2024
Notes

Edit Health Care Agent

Health Care Agents
There are no Health Care Agents on file.

Advance Directive Documents (Scanned)

Document Type	Status	Received On	Received By	Description
Power of Attorney	Not Received			
Advance Directive	Not Received			

Current Code Status

Date Active	Code Status	Order ID	Comments	User	Context
Not on file					

Notes

Date of Service		Author	Author Type	Status	Note Type
12/20/24 1655	Edit		Physician	Cosign Needed	Goals of Care
12/20/24 1439	Edit		Physician	Cosign Needed	Progress Notes