End-to-End Evaluation and Governance of Hyperscribe, an EHR-Embedded Clinical AI Agent

Aaryan Shah[1,2], Andrew Hines[1], Alexia Downs[1], Denis Bajet[1], Paulius Mui[3], Fabiano Araujo[4], Laura Offutt[3], Aida Rutledge[5], Elizabeth Jimenez[3]

[1]Canvas Medical [2]Stanford University, Department of Biomedical Data Science [3]X Primary Care [4]FCA Consulting [5]University of Nevada, Reno, Department of Pediatrics

Continuous governance of deployed clinical AI systems is necessary, achievable, and effective.

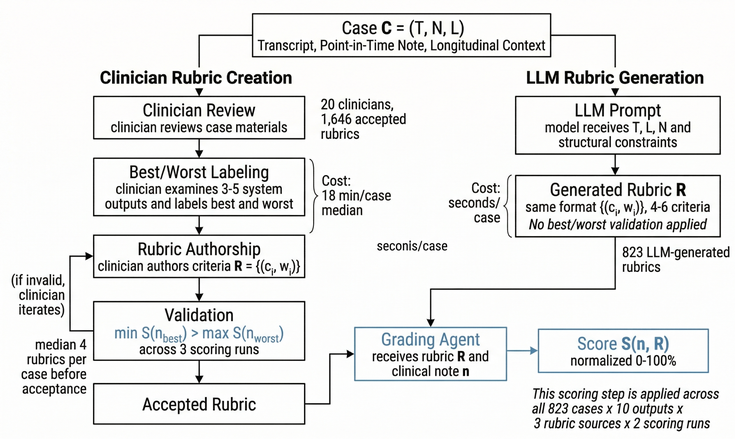
Motivation

Clinical AI systems are rapidly entering practice, yet no published work has demonstrated operational governance of a deployed clinical AI agent with empirical evidence. Governance is the continuous practice of monitoring, evaluating, and iterating clinical AI systems throughout deployment, distinct from point-in-time evaluation.

We present the **first end-to-end governance framework**, validated on a deployed EHR-embedded AI agent for clinicians (Hyperscribe) across Canvas Medical's customer organizations, integrating four evaluation dimensions: **(1) rubric validation** through case-specific rubrics; **(2) live clinician feedback** capturing real-use failures; **(3) technical performance monitoring**; and **(4) cost tracking**. These dimensions are connected by **controlled experimentation** that gates every engineering change before deployment.

Benchmark Construction

Twenty clinicians authored case-specific rubrics through a structured validation platform. Two clinicians reviewed the full transcript and patient context, examined multiple Hyperscribe outputs, labeled best and worst, and authored rubric criteria. Rubrics were validated by confirming that a scoring agent consistently ranked clinician-preferred outputs higher.

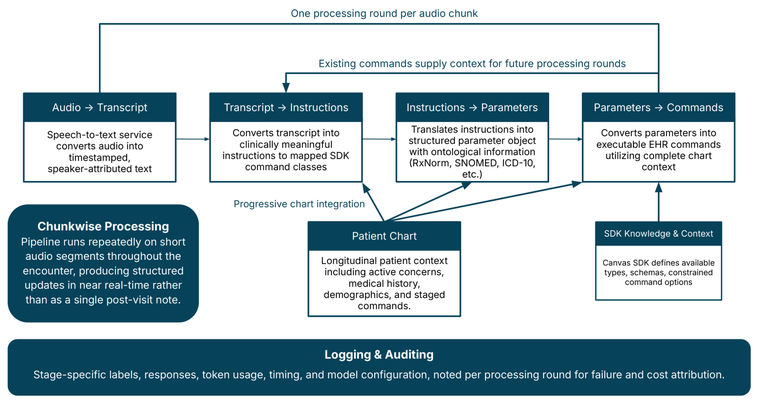


Total cases	823 (736 real, 87 synthetic)
Unique patients	168 across 4 organizations
Accepted rubrics	1,646 from 20 clinicians
Median criteria/rubric	5
Specialties	35 categories

Our dataset is undergoing release through **PhysioNet** under credentialed access, with companion usage scripts on GitHub.

About Hyperscribe

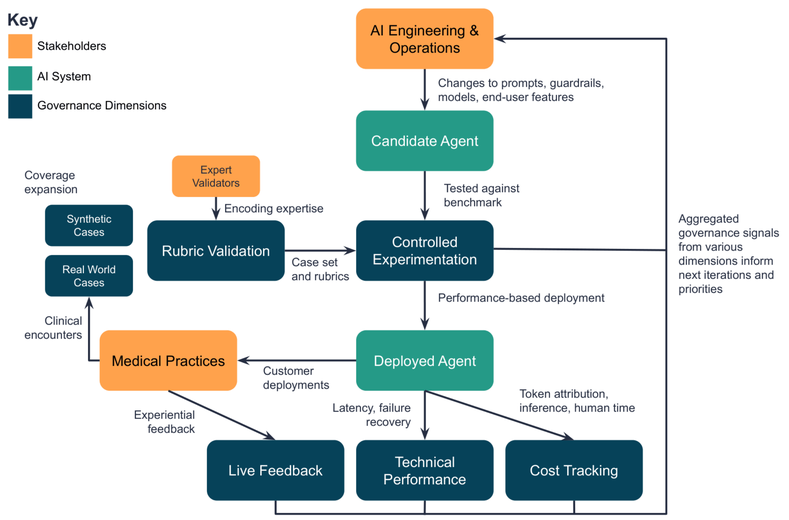
Hyperscribe is an open-source, EHR-embedded ambient scribe developed by Canvas Medical that converts clinical audio to structured chart-updates in near real-time through chunkwise processing.



Hyperscribe was designed to be governable: **(1) structured outputs** produce schema-defined objects; **(2) explicit intermediate reasoning** maintains inspectable layers for failure attribution; **(3) EHR-constrained action space** constrains outputs to predefined clinical actions; and a **(4) computable performance objective** enables rubric-based scoring for version comparison.

Governance Framework

The governance framework integrates four evaluation dimensions with controlled experimentation gating candidate agent versions that are tested against the full benchmark before deployment.



Key Results

Controlled Experimentation. Seven candidate agent versions evaluated through controlled experiments across 823 cases with 10 note simulations per case across model providers:

Versions	Median	Q1	Q3
1–4	83–84%	50–58%	89–90%
5–7	**94–95%**	**80–81%**	**100%**

Feedback-driven Iteration. 107 live feedback entries from Canvas Medical's clinician users over three months of deployment:

Theme	Pre	Post	Change
Command generation	14.0%	3.9%	−10.1
Plan content	10.5%	0.0%	−10.5
Speaker misattribution	3.5%	0.0%	−3.5
Positive observations	14.0%	45.0%	+31.0

Each failure theme led to targeted interventions: prompt rewrites, infrastructure changes, UI redesign, or custom prompting.

Operational Performance. Median processing time **8.1s** per audio segment with an effective completion rate of **99.6%** after retry mechanisms absorbed transient model errors.

Cost of Governance.

Component	Cost	Per unit
Case construction	$320	$0.39/case
Clinician rubrics	919 hrs	17.7 min/rubric
LLM rubrics	$14	$0.02/rubric
Experimentation	$25,000	~$3,600/candidate agent version

Model switching reduced inference costs by 20–30% without quality degradation. Compute costs are marginal, making continuous governance economically sustainable at scale.

Conclusion

Continuous governance of deployed clinical AI is achievable when architectural design, controlled experimentation, and multiple feedback channels operate in concert. Governability must be built in from conception.

LLM-assisted governance makes this feasible: case-specific rubrics at orders of magnitude lower cost enable quality monitoring and assurance previously impractical in clinical settings.

The dataset, source code, evaluation infrastructure, and companion papers are available publicly and at the QR code above.