

MedTrustGraph: Purpose-Aware Temporal Heterogeneous Graph Learning for Zero-Trust Authorization of Autonomous Clinical LLM Agents

KAMRUL ISLAM

kamrulislam01685@gmail.com

Fordham University <https://orcid.org/0000-0001-8291-4970>

SHARMIN SULTANA

Rowan University <https://orcid.org/0009-0000-2445-5438>

Research Article

Keywords: large language models; clinical AI agents; temporal graph neural networks; heterogeneous graphs; Zero Trust; electronic health records; access control; healthcare cybersecurity; FHIR

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11512776/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

MedTrustGraph: Purpose-Aware Temporal Heterogeneous Graph Learning for Zero-Trust Authorization of Autonomous Clinical LLM Agents

SHARMIN SULTANA, Rowan University, USA

KAMRUL ISLAM*, Fordham University, USA

Background: Large language model (LLM) agents can retrieve patient data, invoke clinical tools and execute multi-step tasks inside electronic health record (EHR) systems. Authenticating an agent session does not establish that each subsequent action is consistent with the clinician’s purpose, the assigned patient or the agent’s delegated authority.

Objectives: We study continuous, purpose-aware authorization of individual clinical-agent actions, and test whether temporal graph structure and explicit patient and purpose context improve detection of unauthorized actions over event-level, sequential and static graph models.

Methods: We propose MedTrustGraph, which represents clinicians, agents, patients, tools, FHIR resources and clinical purposes as typed nodes connected by timestamped interactions. A graph attention encoder produces event embeddings that a causal gated recurrent unit processes in temporal order; patient-scope and purpose-scope checks supply Zero-Trust context; and a Platt-calibrated risk estimate is mapped to ALLOW, VERIFY or DENY by a deterministic policy engine that retains final authority. We built a synthetic FHIR-style benchmark modelled on MedAgentBench with 5,000 agent trajectories and 36,855 access events (8,448 unauthorized) spanning cross-patient access, purpose drift, privilege escalation and bulk record retrieval, evaluated on a chronological trajectory-level 70/15/15 split.

Results: Using only leakage-safe observable features, the temporal graph model reached $F1 = 0.8905$, $AUROC = 0.9767$ and $AUPRC = 0.9517$, versus $F1 = 0.8062$ for XGBoost and 0.7918 for a static graph attention network. In a controlled ablation configuration the full model reached $F1 = 0.9653$ (95% CI 0.9544–0.9741); removing patient context, purpose context, graph structure or temporal structure lowered $F1$ by 0.2173, 0.0200, 0.0188 and 0.0174 (all Holm-adjusted $p < 10^{-4}$). Shuffling event order lowered $F1$ to 0.8279, while removing 30% of edges changed it by under 0.002. The calibrated policy layer allowed 0 of 652 unauthorized events (95% CI 0–0.59%) and auto-allowed 98.64% of authorized events with a 1.90% verification workload. The code and data are available at <https://github.com/Kamrul01685/MedTrustGraph>.

Conclusions: On this benchmark, temporal order and patient/purpose context carry authorization signal beyond event-level and static-graph models, and separating learned risk from deterministic policy yields a low-workload decision layer. Because benchmark labels were generated from patient- and purpose-scope rules, context-augmented scores are upper bounds; validation on real or independently generated agent traces is required.

JAIR Associate Editor: Insert JAIR AE Name

JAIR Reference Format:

Sharmin Sultana and Kamrul Islam. 2026. MedTrustGraph: Purpose-Aware Temporal Heterogeneous Graph Learning for Zero-Trust Authorization of Autonomous Clinical LLM Agents. *Journal of Artificial Intelligence Research* 4, Article 6 (August 2026), 24 pages. DOI: [10.1613/jair.1.xxxxx](https://doi.org/10.1613/jair.1.xxxxx)

*Corresponding Author.

Authors’ Contact Information: Sharmin Sultana, ORCID: [0009-0000-2445-5438](https://orcid.org/0009-0000-2445-5438), sultan62@rowan.edu, Rowan University, College of Science & Mathematics, Glassboro, New Jersey, USA; Kamrul Islam, ORCID: [0000-0001-8291-4970](https://orcid.org/0000-0001-8291-4970), ki7@fordham.edu, Fordham University, Information Technology and Operations, New York, New York, USA.

1 Introduction

Large language models are evolving from passive question-answering systems into agents that plan, use external tools, query databases and act in digital environments. In healthcare this matters because clinical workflows require retrieval, longitudinal reasoning, documentation, ordering and coordination across many information sources (W. Wang et al. 2025). LLM agents already operate inside structured EHR environments: MedAgentBench provides a FHIR-compatible virtual EHR with 300 clinically derived tasks, 100 patient profiles and more than 700,000 data elements (Jiang et al. 2025); later work improved agent designs for its multi-step tasks (E. Chen et al. 2026); and autonomous medical agents have been shown to work end to end through clinical cases (Ferber et al. 2026).

Growing autonomy makes authorization harder. A clinician may instruct an agent to review one patient’s laboratory abnormalities, yet during execution the agent may call several tools, retrieve additional resources, delegate work to a sub-agent or respond to malicious content in external data. Medical LLMs are substantially susceptible to prompt injection, including attacks that alter clinically important recommendations (Lee et al. 2025). The authentication event at the start of a session therefore does not establish that every later action is appropriate.

Role-based access control (RBAC) decides whether a physician may use a patient-information system, not whether a particular AI-mediated action at a particular moment is consistent with the clinician’s task. For the instruction “Review Patient P’s recent laboratory results and determine whether renal function has worsened,” access to creatinine, estimated glomerular filtration rate, blood urea nitrogen and relevant medications is justified; access to another patient’s record, bulk extraction of unrelated records or transmission of data to an external service is not, even though the clinician’s credentials are valid.

Zero-Trust Architecture removes implicit trust based on network location or ownership and makes a policy engine responsible for granting, denying or revoking access from policy and context (Chandramouli and Butcher 2023; Rose et al. 2020). The HIPAA Security Rule likewise requires access to electronic protected health information to be limited to authorized users or processes (U.S. Department of Health and Human Services 2026). Zero-Trust designs for healthcare agents address isolation, credentials, network restriction and prompt integrity (Maiti 2026), but not the question: *is the current action justified by the authorized purpose, for this patient, given what the agent has done so far?*

We address this question with *MedTrustGraph*, which models the clinical environment as an evolving heterogeneous graph (Figure 1) and evaluates each agent action in the context of its patient, purpose and behavioural history (Figure 2). Our hypothesis is that unsafe agent behaviour is identified more effectively by jointly modelling relational structure, temporal behaviour, clinical purpose and patient context than by evaluating events independently. The contributions are:

- (1) a formulation of continuous, purpose-aware authorization of clinical-agent actions as event-level risk estimation over a temporal heterogeneous graph;
- (2) a synthetic FHIR-style security benchmark (5,000 trajectories, 36,855 events, four attack families) with chronological trajectory-level splits and an explicit audit of label-leaking features;
- (3) a temporal graph model (graph attention encoder with a causal GRU) and a context-augmented variant, compared with classical, sequential and static graph baselines under identical splits;
- (4) controlled ablation and perturbation analyses with paired trajectory-level bootstrap intervals and exact McNemar tests with Holm correction;
- (5) a calibrated Zero-Trust decision layer in which learned risk is evidence and a deterministic policy engine decides, evaluated on a nested held-out split; and
- (6) a transparent account of the benchmark’s limits, including near-oracle context features and research questions the current data cannot yet answer.

2 Background and Related Work

LLM agents in healthcare. Medical LLM agents add planning, tool use, memory and interaction with clinical systems; applications include decision support, documentation and services, with open challenges in reliability, ethics and safety (W. Wang et al. 2025). MedAgentBench evaluates agents in an interactive FHIR environment across ten task categories, including patient lookup, laboratory retrieval, vital-sign recording, medication ordering and referrals (E. Chen et al. 2026; Jiang et al. 2025).

Security of clinical agents. When a model can read records or call clinical tools, prompt injection, unauthorized retrieval, excessive privilege and data leakage affect real actions rather than only text (Lee et al. 2025). Infrastructure-level Zero-Trust controls for healthcare agents (Maiti 2026) are complementary to our action-level authorization.

Zero-Trust architecture. NIST SP 800-207 bases access decisions on identity, resource attributes, context, policy and observed behaviour, with a Policy Engine that decides and a Policy Administrator and Enforcement Point that implement decisions (Rose et al. 2020). MedTrustGraph follows this separation: the learned model supplies evidence, not enforcement.

FHIR auditability. FHIR supports security labels, provenance, consent and AuditEvent resources that record participants, actions and purpose of use (HL7 International 2023a,b), a natural source of the timestamped interactions from which our graph is built.

Graph and temporal learning. Graph convolutional networks aggregate normalized neighbourhoods (Kipf and Welling 2017); graph attention networks (GAT) learn neighbour-specific weights (Veličković et al. 2018); Heterogeneous Graph Transformers use type-dependent attention (Hu et al. 2020). Temporal Graph Networks learn from event streams with memory modules (Rossi et al. 2020), and recurrent units such as the GRU (Cho et al. 2014) and LSTM (Hochreiter and Schmidhuber 1997) model sequences. Order matters for authorization: ReadLab → ReadMedication → GenerateSummary is expected, whereas ReadLab → Query100Patients → ExportData → ExternalAPI signals escalating risk even when each step is individually permitted.

Learning for security monitoring. Prior JAIR work has modelled security events probabilistically and in continuous time, using continuous-time Bayesian networks for network- and host-based intrusion detection (Xu and Shelton 2010), and has shown how a small amount of labelled data can turn unsupervised anomaly detection into a supervised task (Görnitz et al. 2013). Deep graph representation learning has been surveyed comprehensively, including heterogeneous graphs and the difficulty of dynamic graphs for spectral methods (Chikwendu et al. 2023). MedTrustGraph differs in three ways: it targets the actions of an authenticated AI agent rather than external intrusions, it conditions on an explicit declared intent (purpose and assigned patient), and it places the learned detector under a deterministic policy engine with a three-way decision.

Norms, trust and goal recognition in multi-agent systems. JAIR has a long line of work on keeping autonomous agents accountable. Norm-monitoring mechanisms detect violations of system norms in open multi-agent systems and make monitoring incentive-compatible (Alechina et al. 2018); probabilistic trust models update an agent's trust in others from accumulating evidence as their behaviour changes (Y. Wang et al. 2011); and goal recognition infers what an agent is really trying to achieve, even when it acts deceptively (Le et al. 2021). MedTrustGraph shares their premise that authorization should rest on observed behaviour rather than on identity alone, but it addresses a single tool-using LLM agent whose declared goal (the clinical purpose) is known, asks whether each action is consistent with that goal, and learns the consistency judgement from temporal graph structure rather than specifying it as norms or probabilistic trust updates.

Gap. Prior work shows that agents can act in EHRs, that Zero Trust supports continuous authorization, and that heterogeneous and temporal GNNs model evolving multi-entity interactions. What is missing is a method that asks whether Action_t is justified for Patient_p given Purpose_q , context, policy and behavioural history, rather than merely whether the actor is authenticated.

3 Problem Formulation

3.1 Research questions and hypotheses

We pose six research questions: **RQ1** whether temporal heterogeneous graph modelling improves detection of unauthorized actions over event-level, sequential and static graph models; **RQ2** whether explicit clinical purpose reduces false authorization of contextually inappropriate access; **RQ3** whether patient context improves discrimination between legitimate and unrelated-patient access; **RQ4** whether modelling delegation improves detection of escalation through sub-agents; **RQ5** how well the framework generalizes to unseen patients, tasks, agents and attacks; and **RQ6** whether a policy-grounded LLM can explain decisions faithfully without controlling enforcement. This article answers RQ1–RQ3 empirically, RQ5 partially (chronological hold-out only), and specifies but does not test RQ4 and RQ6 (Section 8).

The hypotheses are: **H1** a temporal GNN outperforms static GNNs on temporally evolving unauthorized behaviour; **H2** adding purpose context to the temporal model improves detection of purpose-inconsistent access; **H3** adding patient context improves performance further; **H4** MedTrustGraph attains a lower false-allow rate than baselines while keeping the false-deny rate acceptable.

3.2 Threat model

The adversary is an authenticated agent acting inside a valid session that a legitimate clinician initiated for a specific patient and purpose. The agent may be misaligned, manipulated by injected content or compromised; we model the actions it attempts, not how it became unsafe. The defender observes each attempted action (actor, tool, resource, action type, target patient, timestamp) together with the session’s assigned patient and declared purpose. The benchmark operationalizes four attack families (Section 5.1). Infrastructure attacks, credential theft, sub-agent delegation, prompt-injection content and network exfiltration are outside the scope of the present evaluation.

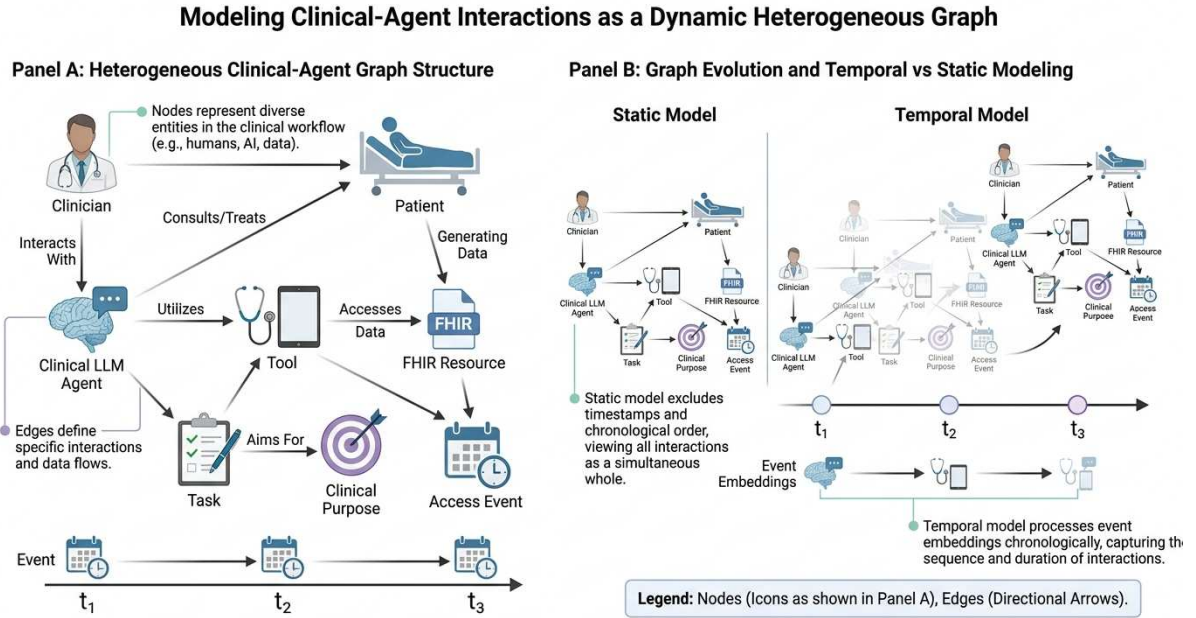
3.3 Task

Let a trajectory τ be a time-ordered sequence of access events $e_1, \dots, e_T$. Each event has a binary label $y_t \in \{0, 1\}$, with $y_t = 1$ denoting an UNAUTHORIZED action. The learning task is to estimate $\Pr(y_t = 1 \mid e_1, \dots, e_t, P_q)$ causally, using only the current and earlier events of the trajectory and the session’s structured purpose P_q , and then to map the estimate to a decision in $\{\text{ALLOW}, \text{VERIFY}, \text{DENY}\}$.

4 The MedTrustGraph Framework

4.1 Overview

MedTrustGraph has six layers: (i) a structured purpose representation of the clinician’s request; (ii) a FHIR/EHR event stream; (iii) a temporal heterogeneous clinical graph; (iv) a temporal graph encoder; (v) context and risk components (patient scope, purpose scope, delegation risk, resource sensitivity) combined into a continuous risk estimate; and (vi) a Zero-Trust policy engine that issues a decision and an evidence-grounded explanation. Figure 1 shows the graph and contrasts static with temporal modelling; Figure 2 contrasts a justified flow with overreach.



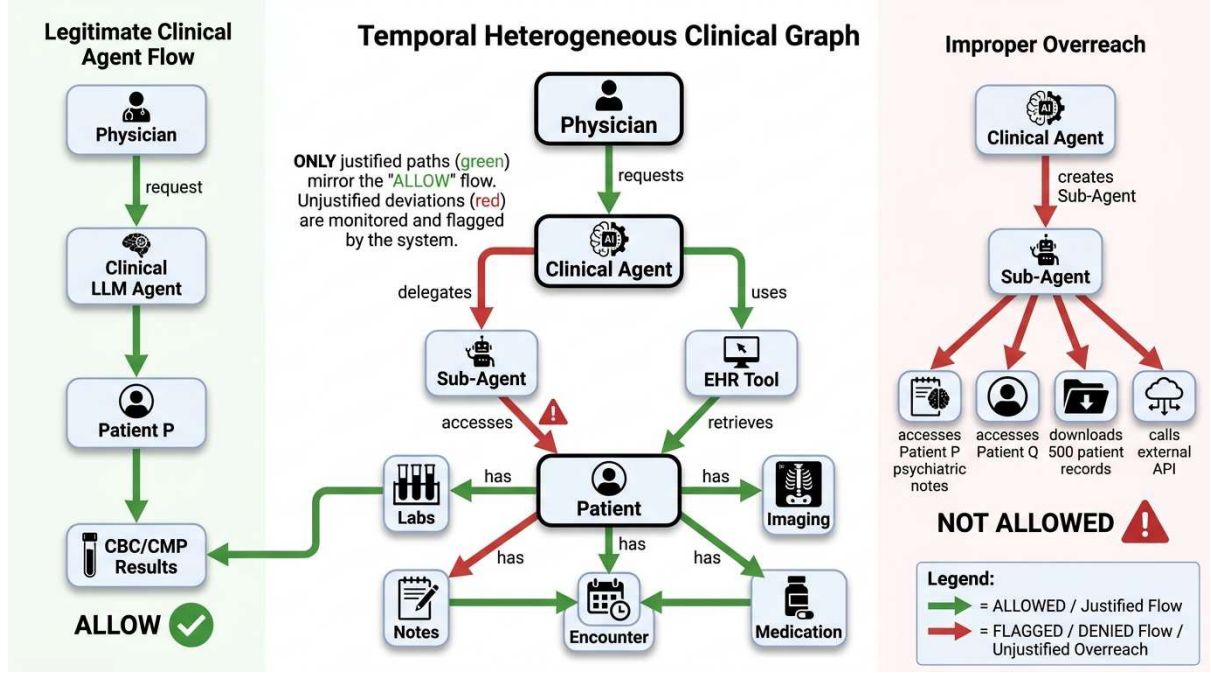


Fig. 2. Justified versus unjustified paths. *Left*: a legitimate flow (physician → agent → assigned patient P → CBC/CMP results) is allowed. *Centre*: justified paths (green) mirror the ALLOW flow; unjustified deviations (red), such as a delegated sub-agent reaching notes outside the task scope, are flagged. *Right*: sub-agent overreach (Patient P’s psychiatric notes, an unrelated Patient Q, a 500-record download, an external API call) is not allowed. Conceptual illustration; delegation, notes, imaging and external-API scenarios belong to the framework design but are absent from the current benchmark.

resource classes = {Observation, MedicationRequest, DiagnosticReport}; concepts = {creatinine, eGFR, BUN, electrolytes, renal medications}; disallowed = {cross-patient retrieval, bulk export, external transmission, deletion}. In deployment an LLM would produce this structure under a fixed schema; in our experiments the purpose is supplied as one of five structured labels, so purpose extraction is not evaluated.

4.4 Zero-Trust context checks

Two context signals are derived from raw relations rather than from any generated label:

$$c_t^{\text{pat}} = \mathbb{1} \left[\text{accessed_patient}_t = \text{assigned_patient}_q \right], \quad (3)$$

$$c_t^{\text{pur}} = \mathbb{1} \left[\text{resource_type}_t \in \mathcal{A}(P_q) \right], \quad (4)$$

where $\mathcal{A}(P_q)$ is the purpose–resource allow-list. In the released implementation, $\mathcal{A}$ is estimated from authorized *training* events only, keeping resource types that account for at least 10% of authorized accesses for each purpose, so no validation or test label is consulted. These signals instantiate a purpose-consistency score $\text{PCS}_t = f(P_q, A_t, R_t, C_t)$ over purpose, action, resource and context.

Algorithm 1 Zero-Trust decision for event t

Input: raw risk p_t ; Platt parameters (a, b) ; thresholds $\theta_1 < \theta_2$; action a_t ; allowed action scope $\mathcal{S}$; context bits $c_t^{\text{pat}}, c_t^{\text{pur}}$

- 1: $\hat{r}_t \leftarrow \sigma(a \cdot \text{logit}(p_t) + b)$
- 2: **if** $\hat{r}_t < \theta_1$ **then** $D \leftarrow \text{ALLOW}$ **else if** $\hat{r}_t < \theta_2$ **then** $D \leftarrow \text{VERIFY}$ **else** $D \leftarrow \text{DENY}$
- 3: **if** $a_t \notin \mathcal{S}$ **then** $D \leftarrow \text{DENY}$ (*least privilege*)
- 4: **if** $D = \text{ALLOW}$ **and** ($c_t^{\text{pat}} = 0$ **or** $c_t^{\text{pur}} = 0$) **then** $D \leftarrow \text{VERIFY}$ (*context cannot auto-allow*)
- 5: **return** D with $\hat{r}_t$ and the triggering rule for audit and explanation

4.5 Temporal graph encoder

Categorical entities receive learned embeddings and continuous attributes are standardized with training statistics. A graph attention encoder (Veličković et al. 2018) aggregates typed neighbours with relation-aware attention,

$$\mathbf{h}_v^{(l+1)} = \sigma \left(\sum_{r \in \mathcal{R}} \sum_{u \in \mathcal{N}_r(v)} \alpha_{uv}^r \mathbf{W}_r \mathbf{h}_u^{(l)} \right), \quad (5)$$

where $\mathcal{R}$ is the set of relation types, $\mathbf{W}_r$ a relation-specific transformation and α_{uv}^r the attention coefficient. For trajectory τ the encoder yields event embeddings $\mathbf{z}_1, \dots, \mathbf{z}_T$, which a causal forward GRU (Cho et al. 2014) processes in timestamp order,

$$\mathbf{s}_t = \text{GRU}(\mathbf{s}_{t-1}, \mathbf{z}_t), \quad (6)$$

so that the state at step t depends only on current and earlier events, mirroring the memory-update principle of Temporal Graph Networks (Rossi et al. 2020).

4.6 Continuous risk estimation

The risk of event t combines the temporal state with the context signals,

$$\text{Risk}_t = \sigma(\text{MLP}([\mathbf{s}_t \parallel c_t^{\text{pat}} \parallel c_t^{\text{pur}}])), \quad T_t = 1 - \text{Risk}_t, \quad (7)$$

where $\parallel$ denotes concatenation and T_t is the trust score; the raw temporal GNN omits the context terms. The framework also defines a delegation risk DR_t as a function of delegation depth, permission inheritance, purpose consistency and agent history, and a resource sensitivity S_r . Sensitivity enters the model as an attribute; delegation risk is not exercised because all benchmark events have delegation depth 0.

4.7 Zero-Trust decision engine

The learned model does not grant access. Its output is calibrated with Platt scaling (Platt 1999) to $\hat{r}_t$, and the policy engine computes

$$D_t = \text{PolicyEngine}(\hat{r}_t, \text{HardPolicy}, \text{Purpose}, \text{Context}) \in \{\text{ALLOW}, \text{VERIFY}, \text{DENY}\}. \quad (8)$$

Thresholds map $\hat{r}_t < \theta_1$ to ALLOW, $\theta_1 \leq \hat{r}_t < \theta_2$ to VERIFY and $\hat{r}_t \geq \theta_2$ to DENY. Deterministic overrides then apply in order: an action outside the least-privilege action scope becomes DENY; a patient-scope or purpose-scope violation that would otherwise be ALLOW becomes VERIFY. Algorithm 1 gives the procedure and Table 1 the decision matrix. Thresholds are selected on a calibration split under safety constraints (Section 5.7).

4.8 Evidence-grounded explanation

For a challenged or denied request, the framework extracts the decisive graph neighbourhood, for example Clinician_7 $\rightarrow$ Agent_12 $\rightarrow$ authorized for Patient_104 (renal assessment) alongside Agent_12 $\rightarrow$ attempted access $\rightarrow$ Patient_208 $\rightarrow$ BehavioralHealthNote. An LLM receives only the purpose, the requested action, this

Table 1. Zero-Trust decision matrix.

Condition	Decision	Rationale
Calibrated risk < 0.45 , no policy violation	ALLOW	Low learned risk; all context checks pass
$0.45 \leq \text{calibrated risk} < 0.92$	VERIFY	Uncertain risk requires human or secondary verification
Calibrated risk ≥ 0.92	DENY	High unauthorized-access risk
Action outside least-privilege scope	DENY	Deterministic least-privilege override
Patient-scope mismatch, otherwise low risk	VERIFY	Cross-patient context cannot be auto-allowed
Purpose-scope mismatch, otherwise low risk	VERIFY	Purpose inconsistency cannot be auto-allowed

evidence, the risk components and the policy identifiers, and may state: “Verification is required because the task is authorized for Patient 104, while the requested behavioural-health record belongs to Patient 208; no clinical relationship supporting cross-patient access was found.” It cannot override a deterministic decision. This component is specified but was not run or evaluated here.

5 Experimental Setup

5.1 Data

Synthetic FHIR records. No real patient data were used. We generated a FHIR R4-style record set for 100 synthetic patients (55 female, 45 male) with 29,600 resources: 20,000 laboratory Observations (creatinine, potassium, magnesium, HbA1c, platelets, sodium, glucose, haemoglobin, blood urea nitrogen, white-cell count), 4,000 MedicationRequests (ten outpatient drugs), 3,000 Conditions (fifteen ICD-10 diagnoses) and 2,500 Procedures (fifteen CPT procedures). The MedAgentBench task set (Jiang et al. 2025) (300 tasks in ten categories of 30) served as a design reference for task types, resource classes and the reference time window; its patients are not linked to the synthetic records.

Security trajectories. From these records we generated 5,000 agent trajectories (seed 20260915). Each is initiated by one of 50 synthetic clinicians, executed by one of 20 agents, assigned to one patient and associated with one of five purposes (anaemia evaluation, diabetes monitoring, hypertension follow-up, medication reconciliation, renal-function review). Agents act through four FHIR tools (Observation, Condition, MedicationRequest, Procedure). Trajectories contain 4–21 events (median 7, mean 7.37), timestamped 1–15 November 2023. There are 3,000 normal trajectories and 500 per attack scenario; attack trajectories begin with authorized actions and therefore mix authorized and unauthorized events:

- *Cross-patient access:* reads of a non-assigned patient (mean 1.95 unauthorized events per trajectory).
- *Purpose drift:* after legitimate actions, reads of a resource class outside the purpose’s scope, e.g. medication requests during anaemia evaluation (mean 3.01).
- *Privilege escalation:* one WRITE, DELETE or EXPORT action in an otherwise read-only trajectory (187, 170 and 143 events).
- *Bulk record retrieval:* a burst of reads across non-assigned patients (mean 10.93; 10-minute access counts up to 21 versus a mean of 3.6 for normal events).

Resource sensitivity is fixed per class (Observation 0.45, Procedure 0.55, MedicationRequest 0.60, Condition 0.65) and delegation depth is 0 for every event. Tables 2 and 3 summarize the data.

Table 2. Benchmark summary.

Component	Value
Synthetic FHIR resources (patients)	29,600 (100)
Trajectories (normal / attack)	5,000 (3,000 / 2,000)
Access events (authorized / unauthorized)	36,855 (28,407 / 8,448; 22.9% unauthorized)
Clinicians / agents / purposes / tools	50 / 20 / 5 / 4
Actions READ / WRITE / DELETE / EXPORT	36,355 / 187 / 170 / 143
Graph nodes	17,274 (17,094 resources, 100 patients, 50 clinicians, 20 agents, 5 tools, 5 purposes)
Graph edges	137,659 (INVOKES, ACCESSES 36,855 each; READS 36,355; HAS_RESOURCE 17,094; ASSIGNS, ACTS_FOR 5,000 each; WRITES 187; DELETES 170; EXPORTS 143)

Table 3. Event-level composition by attack type.

Attack type	Trajectories	Unauthorized events	Mean per trajectory
Normal (28,407 authorized events)	3,000	–	0
Bulk record retrieval	500	5,467	10.93
Purpose drift	500	1,506	3.01
Cross-patient access	500	975	1.95
Privilege escalation	500	500	1.00

Table 4. Chronological trajectory-level split.

Split	Trajectories	Events	Unauthorized	Authorized	Unauthorized rate
Train	3,500	25,781	5,869	19,912	22.8%
Validation	750	5,533	1,253	4,280	22.6%
Test	750	5,541	1,326	4,215	23.9%

5.2 Splitting and leakage control

Random event-level splits leak information across time and within trajectories. Trajectories were sorted by first timestamp and split at the trajectory level into 70% training, 15% validation and 15% test, so no trajectory spans splits and every test trajectory starts after all training trajectories (Table 4). All models use this split.

The generator emits fields that are labels or computed from the labelling rules: `authorization_label`, `attack_type`, `scenario_type`, `is_attack_event`, a heuristic `risk_score`, and the helper fields `patient_match` and `purpose_consistency`. None is a model input in any primary comparison, and timestamps are used only for ordering and splitting. To quantify leakage we also trained classical models on four feature tiers: B0 raw observable features; B1 = B0 + patient context; B2 = B0 + purpose context; B3 = both (Section 6.2).

5.3 Baselines

Classical event-level models: class-balanced logistic regression, random forest (Breiman 2001) with balanced-subsample weights, and XGBoost (T. Chen and Guestrin 2016) with `scale_pos_weight` set to the training class ratio, all using tier B0: resource type, action, purpose, tool and previous action (one-hot) plus resource sensitivity and the 10-minute access count (standardized). *Sequential model:* an LSTM (Hochreiter and Schmidhuber 1997) over the same features in trajectory order. *Static graph models:* a two-layer normalized GCN (Kipf and Welling 2017), a two-layer residual multi-head GAT (Veličković et al. 2018) and an HGT-style transformer (Hu et al. 2020) implemented in PyTorch with node-type-specific query/key/value projections, relation-specific transformations and priors, multi-head attention and residual connections. PyTorch Geometric was unavailable in the execution environment, so this baseline is labelled HGT-style rather than an official HGTConv run. A rule-only RBAC baseline was not evaluated separately: every session is authenticated and every agent may use the EHR tools, so identity-level RBAC alone would allow every event; the least-privilege action rule is evaluated inside the policy layer.

5.4 MedTrustGraph models

Graph construction. Separate event-centred heterogeneous graphs were built per split, with node types Event, Task, Clinician, Agent, Patient, Tool, Resource and Purpose (28-dimensional inputs) and 16 typed relations (event-agent, event-patient, event-tool, event-resource, event-task, task-patient, task-clinician, task-purpose, each with a reverse). The static graphs have 43,629 / 11,105 / 11,155 nodes and 278,810 / 59,830 / 59,910 directed edges for train / validation / test; with temporal edges, 322,439 / 70,935 / 71,065 edges.

Temporal GNN (GAT+GRU). The residual GAT serves as a frozen encoder of event nodes, and a causal forward GRU processes each trajectory in timestamp order to predict authorization at every event, using only leakage-safe observable features (seed 20261003).

MedTrustGraph Full. The temporal GNN plus the two context signals of Section 4.4, derived from raw relations rather than from `patient_match` or `purpose_consistency`.

Controlled ablation configuration. All variants were retrained end to end under one configuration (seed 20261004): Full; minus patient context; minus purpose context; minus graph structure (relation attention replaced by mean pooling of entity embeddings); minus temporal structure (GRU removed); and a raw temporal model without context. Absolute scores in this configuration are compared within the ablation table.

Training. Neural models were selected by validation F1 with early stopping. The released reference implementation trains with class-weighted binary cross-entropy and AdamW (Loshchilov and Hutter 2019), 24-dimensional embeddings, 64-dimensional hidden layers, learning rate 10^{-3} , weight decay 10^{-4} , at most 30 epochs and patience 6 (Appendix A). The classification threshold is 0.5.

5.5 Metrics

Unauthorized events are the minority class, so we report precision, recall, F1, AUROC and AUPRC, the last being informative under imbalance (Saito and Rehmsmeier 2015). With unauthorized as the positive class, the false-allow rate is $FAR = FN / (FN + TP)$ and the false-deny rate is $FDR = FP / (FP + TN)$. We emphasize FAR because a missed unauthorized action is the costliest error in this setting. Calibration is measured by the Brier score (Brier 1950) and 10-bin expected calibration error (ECE) (Guo et al. 2017). For the policy layer we report decision proportions per true class, the security capture rate (VERIFY+DENY among unauthorized events), the VERIFY workload and the auto-resolution rate.

5.6 Robustness protocol

The trained Full model was evaluated under test-time perturbations only: random removal of 10%, 20% and 30% of typed edges; random permutation of event order within each trajectory; and random flipping of 5%, 10% and 20% of context bits.

5.7 Calibration and policy protocol

The 750 test trajectories of the controlled Full model were split chronologically into halves. The first (2,754 events, 674 unauthorized) was used to fit Platt scaling and to grid-search thresholds subject to unauthorized $\text{ALLOW} \leq 1\%$ and authorized $\text{DENY} \leq 3\%$, minimizing a security-weighted cost. Six settings tied at minimum cost ($\theta_1 \in [0.45, 0.50]$, $\theta_2 = 0.92$); the most conservative, $\theta_1 = 0.45$, $\theta_2 = 0.92$, was locked. The second half (2,787 events, 652 unauthorized) was evaluated once. Because these trajectories had been inspected in earlier experiments, this is a nested secondary evaluation, not an untouched external cohort. All normal trajectories are read-only, so the least-privilege scope is $\{\text{READ}\}$; a deployment must take scopes from institutional RBAC/ABAC and delegation policy.

5.8 Statistical analysis

F1 confidence intervals used trajectory-level bootstrap resampling (Efron and Tibshirani 1993) with 500 replicates. Differences between the Full model and each ablation used a paired trajectory-level bootstrap with 5,000 replicates, and paired event-level correctness was compared with exact McNemar tests (McNemar 1947) with Holm correction over five comparisons (Holm 1979). Wilson intervals (Wilson 1927) were computed for policy rates. Each model was trained once (one run per reported result), so comparisons among separately trained baselines are descriptive.

6 Results

6.1 Main benchmark (RQ1)

Table 5 reports test performance. On raw observable features, XGBoost was the strongest conventional baseline (F1 0.8062, AUPRC 0.9134) and GAT the strongest static graph model (F1 0.7918, AUPRC 0.8691). Static structure alone did not beat strong tabular or sequential baselines: GCN (F1 0.5292) and HGT-style (F1 0.5847) lost signal through neighbourhood aggregation. The temporal GNN reached F1 0.8905, AUROC 0.9767 and AUPRC 0.9517, with precision 0.9287 and FDR 0.0206. Relative to XGBoost it gained 0.0842 F1 (10.4% relative) and 0.0383 AUPRC; relative to GAT, 0.0987 F1 (12.5%) and 0.0826 AUPRC. Its FAR (0.1448) was close to the best baselines (0.1486–0.1606); most of the gain came from fewer false denials (87 versus 322 for XGBoost). Figure 3 shows the progression. Training took 10.5 s (20 epochs; best validation F1 0.8855) for the temporal GNN and 11.5 s (21 epochs) for the context-augmented model; inference latency and memory were not measured.

6.2 Context features as near-oracles

Adding context made the synthetic task nearly trivial for classical models (Table 6). With patient context alone, random forest and XGBoost reached perfect test scores and logistic regression F1 0.9346; with purpose context every model reached F1 1.0, as did the context-augmented temporal GNN. The generator labels cross-patient and bulk events by patient mismatch and purpose-drift events by out-of-scope resource classes, so the context signals nearly restate the labelling rules. We therefore treat the raw temporal GNN as the fair learned-model comparison and context-augmented results as upper bounds that validate the policy-context pathway, not as evidence of real-world performance.

Table 5. Test performance (750 trajectories, 5,541 events, threshold 0.5). FAR: false-allow rate; FDR: false-deny rate. [†]Upper bound: context features mirror the label-generation rules (Section 6.2). [‡]Separate controlled retraining used for ablations; not directly comparable with the rows above.

Model	Acc.	Prec.	Rec.	F1	AUROC	AUPRC	FAR	FDR
Logistic regression	.8717	.6921	.8356	.7571	.9486	.8831	.1644	.1170
Random forest	.8973	.7525	.8507	.7986	.9596	.9132	.1493	.0880
XGBoost	.9034	.7756	.8394	.8062	.9597	.9134	.1606	.0764
LSTM	.8959	.7482	.8514	.7965	.9577	.9081	.1486	.0902
GCN	.7342	.4592	.6244	.5292	.7304	.4195	.3756	.2313
GAT	.8957	.7579	.8288	.7918	.9376	.8691	.1712	.0833
HGT-style	.7526	.4886	.7278	.5847	.7782	.4977	.2722	.2396
Temporal GNN (GAT+GRU)	.9496	.9287	.8552	.8905	.9767	.9517	.1448	.0206
MedTrustGraph Full [†]	1.000	1.000	1.000	1.000	1.000	1.000	.0000	.0000
Full MedTrustGraph (controlled) [‡]	.9836	.9776	.9532	.9653	.9984	.9952	.0468	.0069

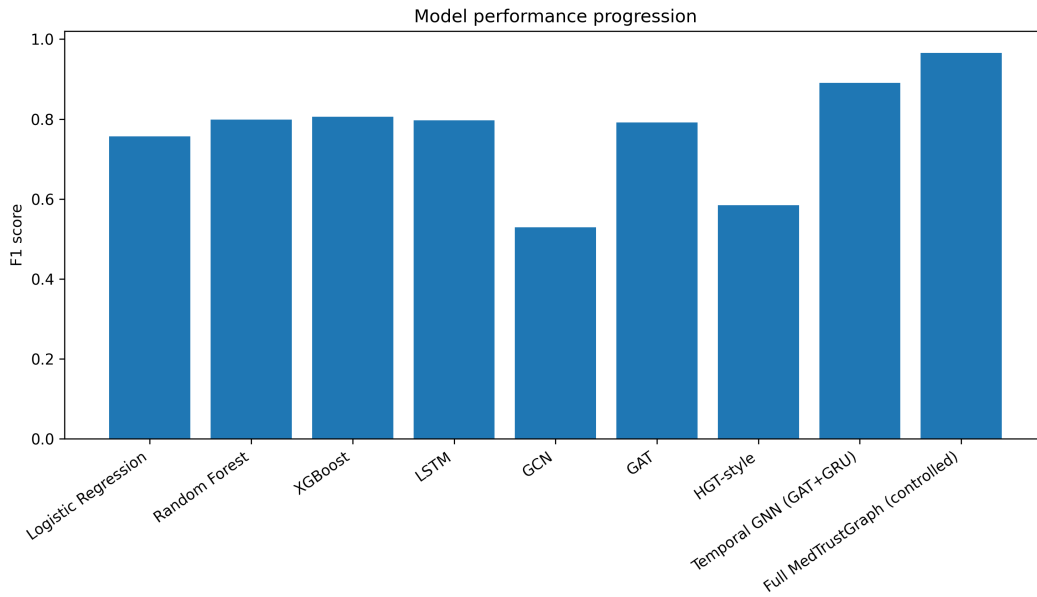


Fig. 3. Test F1 across model families. The final bar is the controlled Full configuration, trained separately for ablation analysis.

6.3 Ablation analysis (RQ2, RQ3)

In the controlled configuration the Full model reached F1 0.9653 (95% CI 0.9544–0.9741), FAR 0.0468 and FDR 0.0069 (Table 7). Removing patient context caused the largest loss (F1 0.7480), driven by cross-patient recall falling from 0.9936 to 0.5064 and authorized specificity from 0.9931 to 0.8797. Removing purpose context, graph structure or temporal structure produced smaller but consistent losses. All five paired comparisons were significant after Holm correction, with bootstrap probability 1.0 that the Full model was better in each case (Table 8; Figure 4).

Table 6. Test F1 of classical models by feature tier.

Feature tier	Logistic regression	Random forest	XGBoost
B0 Raw observable	0.7571	0.7986	0.8062
B1 + Patient context	0.9346	1.0000	1.0000
B2 + Purpose context	1.0000	1.0000	1.0000
B3 Full context	1.0000	1.0000	1.0000

Table 7. Controlled ablation results (test set). F1 intervals from 500 trajectory-level bootstrap replicates.

Configuration	Prec.	Rec.	F1 [95% CI]	AUROC	AUPRC	FAR	FDR	$\Delta F1$
Full model	.9776	.9532	.9653 [.9544, .9741]	.9984	.9952	.0468	.0069	–
– Patient context	.6835	.8258	.7480 [.7141, .7760]	.9399	.8593	.1742	.1203	–.2173
– Purpose context	.9661	.9253	.9453 [.9306, .9566]	.9937	.9859	.0747	.0102	–.0200
– Graph structure	.9540	.9389	.9464 [.9312, .9598]	.9950	.9873	.0611	.0142	–.0188
– Temporal struct.	.9570	.9389	.9478 [.9349, .9597]	.9971	.9918	.0611	.0133	–.0174
Raw temporal (no context)	.7377	.8039	.7694 [.7392, .7989]	.9461	.8764	.1961	.0899	–.1959

Table 8. Paired comparisons of the Full model with each ablation: paired trajectory bootstrap (5,000 replicates) and exact McNemar test on 5,541 paired events, Holm-corrected over five comparisons. n_{10} : Full correct, ablation wrong; n_{01} : the reverse.

Comparison	$\Delta F1$	95% CI	n_{10}	n_{01}	Holm p
vs – Patient context	0.2173	[0.1901, 0.2485]	684	37	1.5×10^{-154}
vs – Purpose context	0.0200	[0.0131, 0.0278]	64	13	9.1×10^{-9}
vs – Graph structure	0.0188	[0.0108, 0.0278]	69	19	1.6×10^{-7}
vs – Temporal struct.	0.0174	[0.0088, 0.0267]	81	35	2.3×10^{-5}
vs Raw temporal	0.1959	[0.1714, 0.2238]	579	31	2.5×10^{-131}

6.4 Robustness

Removing up to 30% of test edges changed F1 by at most 0.0019 (Table 9; Figure 5), indicating redundancy in static relational structure. Permuting event order within trajectories lowered F1 from 0.9653 to 0.8279, direct evidence within the benchmark that order carries authorization signal. Context corruption degraded F1 progressively (0.9240 at 5%, 0.8870 at 10%, 0.8174 at 20%) while FAR stayed between 0.0377 and 0.0513, so losses came mostly from extra false denials; the system is sensitive to the quality of its context inputs.

6.5 Attack-specific detection

Table 10 gives recall per attack family and specificity on authorized events. The Full model detected bulk retrieval perfectly and cross-patient access almost perfectly (0.9936), with lower recall on purpose drift (0.8745) and privilege escalation (0.6076) (Figure 6). Patient context was essential for cross-patient access (0.5064 without it). Privilege escalation depended on the full combination of components: recall fell to 0.0759 without purpose context and 0.2785 without temporal structure, while the raw temporal model reached 0.7468. Because each such

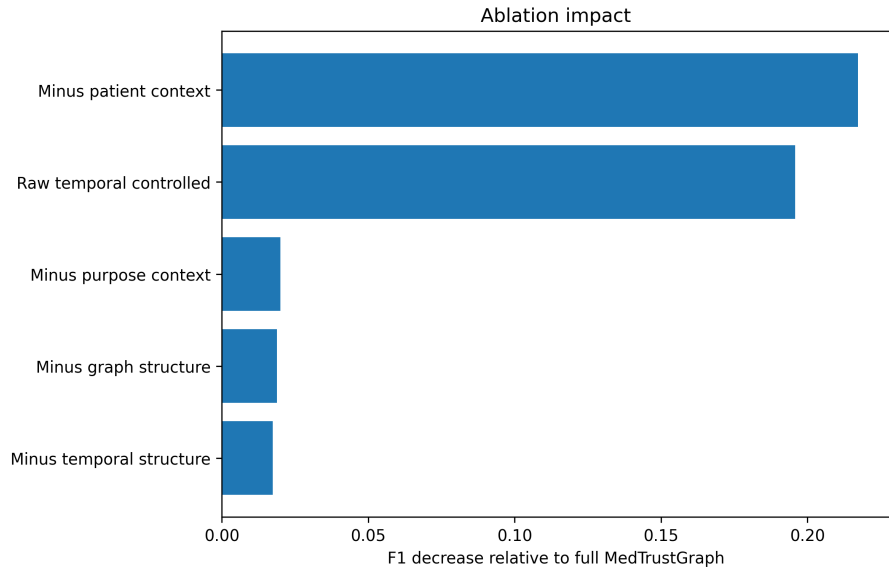


Fig. 4. Decrease in test F1 relative to the Full model when each component is removed (controlled configuration).

Table 9. Robustness of the Full model to test-time perturbations.

Perturbation	Level	F1	AUROC	FAR	$\Delta F1$
None (clean)	0	0.9653	0.9984	0.0468	–
Random edge drop	10%	0.9641	0.9983	0.0475	–0.0012
Random edge drop	20%	0.9635	0.9982	0.0452	–0.0018
Random edge drop	30%	0.9634	0.9979	0.0468	–0.0019
Shuffled event order	all	0.8279	0.9713	0.0641	–0.1374
Context-bit corruption	5%	0.9240	0.9937	0.0513	–0.0413
Context-bit corruption	10%	0.8870	0.9877	0.0437	–0.0783
Context-bit corruption	20%	0.8174	0.9784	0.0377	–0.1479

trajectory contains a single non-READ action, action scope is better enforced by an explicit permission rule than learned through the context pathway, which motivated the least-privilege override.

6.6 Zero-Trust decision layer (H4)

On the locked evaluation half no unauthorized event was allowed (0/652; 95% Wilson CI 0.00–0.59%); 94.94% were denied and 5.06% verified, a security capture rate of 100%. Authorized events were allowed in 98.64% of cases, verified in 0.94% and denied in 0.42%; the overall VERIFY workload was 1.90% and 98.10% of events were resolved automatically (Table 11). The calibration half gave closely matching rates (unauthorized ALLOW 0.00%, authorized ALLOW 99.18%, workload 1.74%). Every attack family was fully captured (Table 12; Figure 7); purpose drift had the largest share sent to verification (22.8%), consistent with plausible but out-of-scope reads. The least-privilege

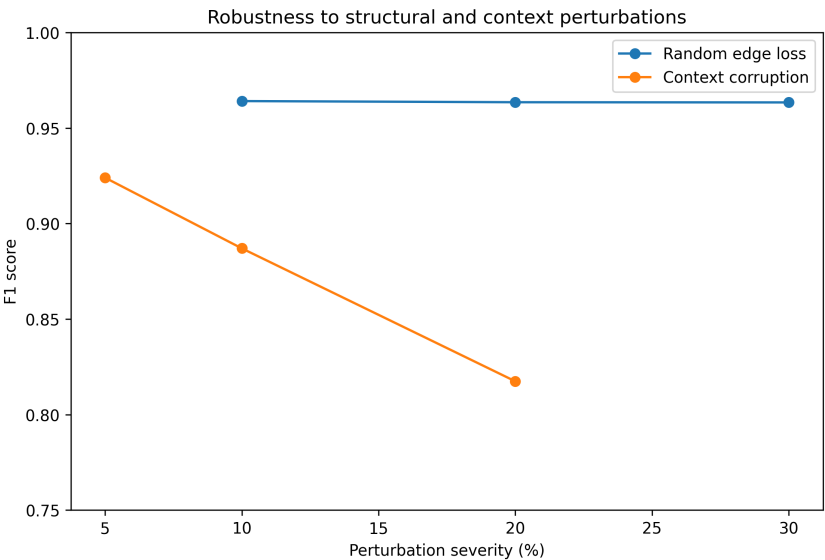


Fig. 5. Full-model F1 under random edge loss and context-bit corruption; event-order shuffling (F1 0.8279) is reported in Table 9.

Table 10. Attack-wise recall and authorized-event specificity (controlled configuration).

Configuration	Bulk (n=852)	Cross-patient (n=156)	Purpose drift (n=239)	Privilege esc. (n=79)	Authorized spec. (n=4,215)
Full MedTrustGraph	1.0000	0.9936	0.8745	0.6076	0.9931
– Patient context	0.8791	0.5064	0.9791	0.4177	0.8797
– Purpose context	1.0000	1.0000	0.8912	0.0759	0.9898
– Graph structure	1.0000	1.0000	0.7824	0.6329	0.9858
– Temporal struct.	1.0000	1.0000	0.8996	0.2785	0.9867
Raw temporal (no context)	0.8650	0.4295	0.8494	0.7468	0.9101

override was decisive for privilege escalation: without it, 4 of 35 such events (11.4%) would have been allowed and 68.6% verified; with it, all 35 were denied.

6.7 Calibration

Platt scaling preserved discrimination (AUROC 0.9984, AUPRC 0.9947 on the evaluation half) and reduced the Brier score from 0.0184 to 0.0121 (34.4% relative) and ECE from 0.0692 to 0.0058 (91.6%) (Table 13). Calibrated risks concentrated at the extremes: 2,093 evaluation events fell in the lowest bin (mean risk 0.0049, observed rate 0.0038) and 609 in the highest (0.9946 versus 0.9803). Intermediate bins held 4–27 events each with larger gaps (Figure 8), so calibration in the VERIFY range is less precisely estimated.

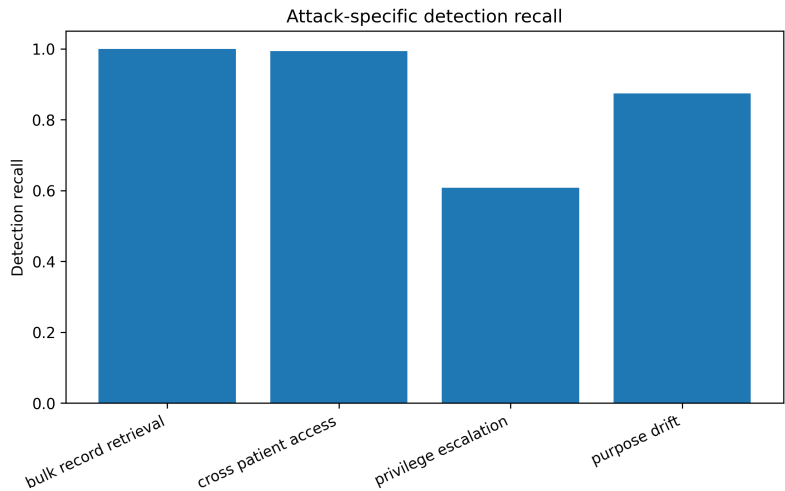


Fig. 6. Attack-specific detection recall of the Full model.

Table 11. Policy decisions on the evaluation half with 95% Wilson intervals.

True class	Decision	Count / <i>n</i>	Rate	95% CI
Unauthorized	ALLOW	0 / 652	0.00%	0.00–0.59%
Unauthorized	VERIFY	33 / 652	5.06%	3.63–7.02%
Unauthorized	DENY	619 / 652	94.94%	92.98–96.37%
Authorized	ALLOW	2,106 / 2,135	98.64%	98.06–99.05%
Authorized	VERIFY	20 / 2,135	0.94%	0.61–1.44%
Authorized	DENY	9 / 2,135	0.42%	0.22–0.80%

Table 12. Policy decisions by attack type (evaluation half).

Attack type	Events	DENY	VERIFY	ALLOW	Capture
Bulk record retrieval	426	99.77%	0.23%	0.00%	100%
Cross-patient access	68	94.12%	5.88%	0.00%	100%
Privilege escalation	35	100.00%	0.00%	0.00%	100%
Purpose drift	123	77.24%	22.76%	0.00%	100%
Normal authorized	2,135	0.42%	0.94%	98.64%	–

7 Discussion

Research questions. RQ1 is supported descriptively: on observable features the temporal GNN exceeded every baseline on F1, AUROC and AUPRC. Two results point to order as the source of the gain: static structure alone did not beat tabular models, and shuffling order removed 0.1374 F1 while removing 30% of edges removed under 0.002. RQ2 is supported in the controlled configuration: removing purpose context lowered F1 by 0.0200 and raised FAR from 0.0468 to 0.0747; the effect is modest because patient context already captures cross-patient and

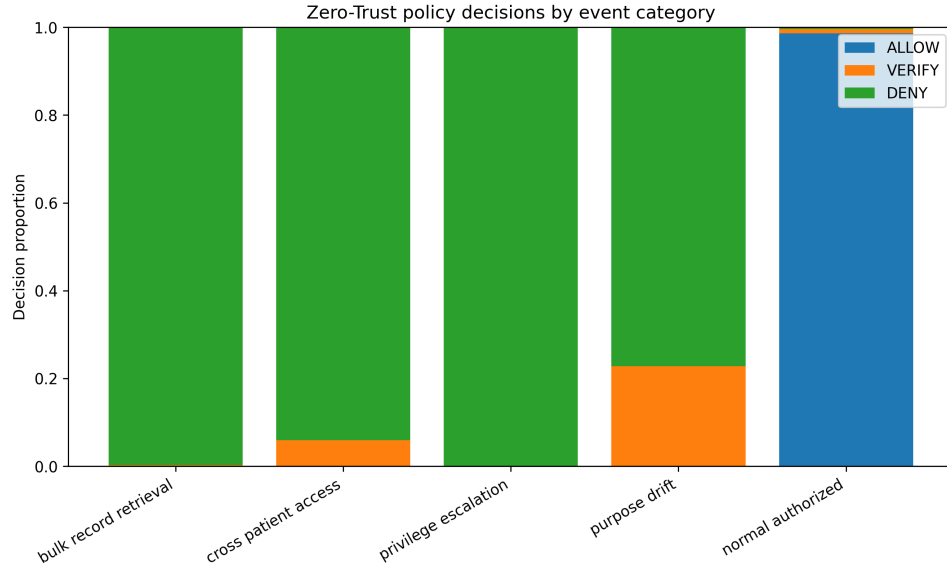


Fig. 7. Proportion of ALLOW, VERIFY and DENY decisions per event category on the evaluation half.

Table 13. Calibration before and after Platt scaling.

Split	Score	Brier	ECE	AUROC	AUPRC
Calibration half	Raw	0.0183	0.0728	0.9984	0.9957
Calibration half	Platt	0.0112	0.0054	0.9984	0.9957
Evaluation half	Raw	0.0184	0.0692	0.9984	0.9947
Evaluation half	Platt	0.0121	0.0058	0.9984	0.9947

bulk attacks. *RQ3* is supported with the largest effect ($\Delta F1 = 0.2173$), partly reflecting that patient mismatch is one of the attack-generation rules. *RQ4* was not tested (delegation depth is always 0). *RQ5* is partially addressed: test trajectories are strictly later, but patients, agents, tools and attack families overlap across splits. *RQ6* was not tested.

Hypotheses. Table 14 summarizes the evidence. All four hypotheses are supported on the synthetic benchmark, with H1 supported descriptively and H3 dependent on how the benchmark was generated.

Separating prediction from authorization. The classifier alone missed 39% of privilege-escalation events, yet a single deterministic least-privilege rule eliminated all of them, and the scope overrides ensured no context violation was auto-allowed. Conversely, learned risk provides what rules cannot: graded risk for sequences such as purpose drift, where 22.8% of attack events were verified rather than denied. A three-way decision with a small VERIFY band (1.90%) keeps human workload low while avoiding forced binary outcomes on uncertain cases, mirroring the NIST division between a policy engine that weighs evidence and the components that enforce it (Rose et al. 2020).

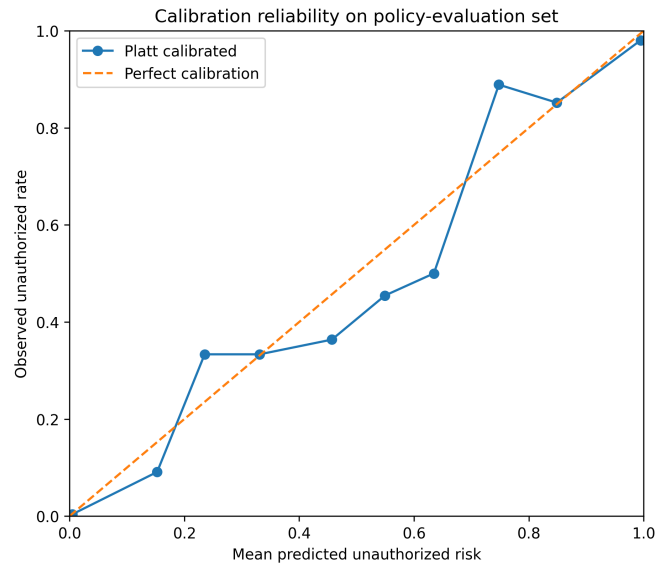


Fig. 8. Reliability diagram of Platt-calibrated risk on the evaluation half; the dashed line is perfect calibration.

Table 14. Status of the hypotheses on the synthetic benchmark.

Hypothesis	Evidence	Status
H1: temporal > static GNN	F1 0.8905 vs 0.7918 (GAT); AUPRC 0.9517 vs 0.8691	Supported (descriptive, one run)
H2: +purpose helps	Full vs −purpose: $\Delta F1$ 0.0200 [0.0131, 0.0278], Holm $p = 9.1 \times 10^{-9}$	Supported (controlled)
H3: +patient helps further	Full vs −patient: $\Delta F1$ 0.2173 [0.1901, 0.2485], Holm $p < 10^{-150}$	Supported (benchmark-dependent)
H4: lower FAR, acceptable FDR	Controlled Full FAR 0.0468 vs 0.1448–0.3756 for raw-feature models, FDR 0.0069; policy layer 0/652 unauthorized allowed, 0.42% authorized denied	Supported (synthetic)

Significance for AI. Beyond healthcare, the setting is an instance of a general problem in agentic AI: authorizing individual actions of a tool-using agent against a declared intent, under distribution shift and adversarial pressure. Our results suggest three transferable lessons: model behaviour as an ordered event stream rather than as independent requests; make intent and scope explicit and machine-checkable; and keep learned components advisory, calibrated and subordinate to auditable rules.

Clinical and security relevance. As agents take on documentation, summarization, laboratory review, medication reconciliation, coding, prior authorization, messaging and scheduling, security controls must decide whether each action is needed for the current task, not only whether the agent can reach the EHR. MedTrustGraph is designed as an additional contextual authorization layer between an agent’s request and the policy engine, complementing

the access controls, authentication, auditing and transmission security required under HIPAA (U.S. Department of Health and Human Services 2026).

8 Limitations

- (1) **Synthetic data.** All records and trajectories are synthetic; they are neither real patient data nor the MedAgentBench FHIR database, and a virtual environment cannot reproduce a production hospital.
- (2) **Rule-generated labels.** Attacks were generated from rules tied to patient scope, purpose–resource compatibility, action type and access volume, so context features act as near-oracles (Table 6) and context-augmented scores are upper bounds.
- (3) **Threat coverage.** Only four attack families are present. Delegation, prompt-injection content, exfiltration, sensitive-record classes and adaptive adversaries are absent, so RQ4 and parts of Figure 2 are untested.
- (4) **Evaluation design.** Each model was trained once, with seeds differing across experimental stages; baseline comparisons are descriptive. The policy layer was evaluated on a nested half of a previously inspected test set.
- (5) **Implementation.** The HGT baseline is a PyTorch HGT-style implementation, not HGTConv. The consolidated reference code reproduces the methodology but not necessarily every reported figure across hardware and library versions.
- (6) **Benchmark-specific policy.** The read-only action scope reflects the benchmark; deployments need scopes from RBAC/ABAC, delegation and care context.
- (7) **Unevaluated components.** LLM purpose extraction, explanation generation, time-to-detection, latency and memory were not measured.
- (8) **Clinical trade-offs and transfer.** Legitimate care often needs information that looks only indirectly related to the stated question, and over-restrictive authorization could delay care. Policies differ across institutions, so models may not transfer between sites. The framework addresses authorization, not clinical decision-making.

9 Future Work

Priorities are: evaluation on traces produced by real LLM agents in MedAgentBench-style environments, with attacks generated independently of the feature definitions and with patient-, agent-, tool- and attack-held-out splits; delegation chains, prompt-injection-induced tool calls, exfiltration and sensitive-record classes with explicit permission inheritance; implementation and evaluation of the explainer (policy attribution accuracy, evidence faithfulness, hallucination rate, consistency); counterfactual authorization that identifies the minimum change making a denied action permissible; federated and continual learning across institutions; multimodal graphs and agent-to-agent trust propagation; and human–agent collaborative authorization that decides when to escalate rather than deny.

10 Conclusion

MedTrustGraph frames authorization of autonomous clinical LLM agents as continuous, purpose-aware risk estimation over a temporal heterogeneous graph, with a deterministic Zero-Trust policy engine holding final authority. On a synthetic FHIR-style benchmark, a temporal graph model on observable features outperformed event-level, sequential and static graph baselines (F1 0.8905 versus 0.8062 and 0.7918); controlled ablations showed significant contributions from patient context, purpose context, graph structure and temporal structure; and the calibrated policy layer allowed none of the evaluated unauthorized events while auto-allowing 98.64% of authorized ones with a 1.90% verification workload. These results validate the design and pipeline rather than real-world performance. If they hold on realistic agent traces, the approach can support architectures in which

clinical AI agents receive neither permanent trust nor unrestricted access, but must continuously show that each action is justified by patient, purpose, policy and context.

11 Data, Code and Ethics

11.1 Data and code availability

The synthetic FHIR-style records, the security-event dataset (events, trajectory summaries, graph nodes and edges, data dictionary and manifest), split assignments, model checkpoints, per-event predictions and the reference implementation are publicly available at <https://github.com/Kamrul01685/MedTrustGraph>. They will be released under an open licence (MIT for code, CC BY 4.0 for data) and also provided as a JAIR online appendix upon publication.

11.2 Ethics statement

All experiments used synthetic data in a sandboxed environment; no identifiable patient information was used or sent to an external LLM service. Deployment should process identifiable data only under appropriate legal, privacy and institutional frameworks, keep complete audit trails of actions, requests, risk scores, decisions, explanations and human overrides, and retain human review for VERIFY decisions.

References

- N. Alechina, J. Y. Halpern, I. A. Kash, and B. Logan. 2018. "Incentive-Compatible Mechanisms for Norm Monitoring in Open Multi-Agent Systems." *Journal of Artificial Intelligence Research*, 62, 433–458. doi:10.1613/jair.1.11214.
- L. Breiman. 2001. "Random Forests." *Machine Learning*, 45, 1, 5–32.
- G. W. Brier. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, 78, 1, 1–3.
- R. Chandramouli and Z. Butcher. 2023. *A Zero Trust Architecture Model for Access Control in Cloud-Native Applications in Multi-Cloud Environments*. NIST Special Publication 800-207A. National Institute of Standards and Technology.
- E. Chen, S. Postelnik, K. Black, Y. Jiang, and J. H. Chen. 2026. "MedAgentBench v2: Improving Medical LLM Agent Design." In: *Pacific Symposium on Biocomputing*. Vol. 31, 354–371.
- T. Chen and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- I. A. Chikwendu, X. Zhang, I. O. Agyemang, I. Adjei-Mensah, U. C. Chima, and C. J. Ejayi. 2023. "A Comprehensive Survey on Deep Graph Representation Learning Methods." *Journal of Artificial Intelligence Research*, 78, 287–356.
- K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1724–1734.
- B. Efron and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall.
- D. Ferber et al.. 2026. "Towards Autonomous Medical Artificial Intelligence Agents." *Nature*, 655, 1282–1291. doi:10.1038/s41586-026-10675-5.
- N. Görmitz, M. Kloft, K. Rieck, and U. Brefeld. 2013. "Toward Supervised Anomaly Detection." *Journal of Artificial Intelligence Research*, 46, 235–262.
- O. E. Gundersen et al.. 2024. "Improving Reproducibility in AI Research: Four Mechanisms Adopted by JAIR." *Journal of Artificial Intelligence Research*, 81, 1019–1041.
- C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. 2017. "On Calibration of Modern Neural Networks." In: *Proceedings of the 34th International Conference on Machine Learning (PMLR)*. Vol. 70, 1321–1330.
- HL7 International. 2023a. *FHIR AuditEvent Resource*. FHIR Release 5. (2023).
- HL7 International. 2023b. *FHIR Release 5: Security and Privacy Module*. Fast Healthcare Interoperability Resources Specification. (2023).
- S. Hochreiter and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation*, 9, 8, 1735–1780.
- S. Holm. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*, 6, 2, 65–70.
- Z. Hu, Y. Dong, K. Wang, and Y. Sun. 2020. "Heterogeneous Graph Transformer." In: *Proceedings of The Web Conference 2020*, 2704–2710. doi:10.1145/3366423.3380027.
- Y. Jiang, K. C. Black, G. Geng, D. Park, J. Zou, A. Y. Ng, and J. H. Chen. 2025. "MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents." *NEJM AI*, 2, 9. doi:10.1056/AIdbp2500144.

- T. N. Kipf and M. Welling. 2017. "Semi-Supervised Classification with Graph Convolutional Networks." In: *International Conference on Learning Representations*.
- T. Le, R. Singh, and T. Miller. 2021. "Goal Recognition for Deceptive Human Agents through Planning and Gaze." *Journal of Artificial Intelligence Research*, 71, 697–732. doi:10.1613/jair.1.12518.
- R. W. Lee, T. J. Jun, J. Lee, S. I. Cho, H. J. Park, and J. Suh. 2025. "Vulnerability of Large Language Models to Prompt Injection When Providing Medical Advice." *JAMA Network Open*, 8, 12, e2549963. doi:10.1001/jamanetworkopen.2025.49963.
- I. Loshchilov and F. Hutter. 2019. "Decoupled Weight Decay Regularization." In: *International Conference on Learning Representations*.
- S. Maiti. 2026. *Caging the Agents: A Zero Trust Security Architecture for Autonomous AI in Healthcare*. arXiv:2603.17419. (2026).
- Q. McNemar. 1947. "Note on the Sampling Error of the Difference between Correlated Proportions or Percentages." *Psychometrika*, 12, 2, 153–157.
- J. Platt. 1999. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods." In: *Advances in Large Margin Classifiers*. MIT Press, 61–74.
- S. Rose, O. Borchert, S. Mitchell, and S. Connelly. 2020. *Zero Trust Architecture*. NIST Special Publication 800-207. National Institute of Standards and Technology. doi:10.6028/NIST.SP.800-207.
- E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein. 2020. "Temporal Graph Networks for Deep Learning on Dynamic Graphs." In: *ICML 2020 Workshop on Graph Representation Learning and Beyond*. arXiv:2006.10637.
- T. Saito and M. Rehmsmeier. 2015. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLoS ONE*, 10, 3, e0118432.
- U.S. Department of Health and Human Services. 2026. *Summary of the HIPAA Security Rule*. Office for Civil Rights. (2026).
- P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. 2018. "Graph Attention Networks." In: *International Conference on Learning Representations*.
- W. Wang, Z. Ma, Z. Wang, et al.. 2025. "A Survey of LLM-based Agents in Medicine: How far are we from Baymax?" In: *Findings of the Association for Computational Linguistics: ACL 2025*, 10345–10359. doi:10.18653/v1/2025.findings-acl.539.
- Y. Wang, C. Hang, and M. P. Singh. 2011. "A Probabilistic Approach for Maintaining Trust Based on Evidence." *Journal of Artificial Intelligence Research*, 40, 221–267. doi:10.1613/jair.3108.
- E. B. Wilson. 1927. "Probable Inference, the Law of Succession, and Statistical Inference." *Journal of the American Statistical Association*, 22, 158, 209–212.
- J. Xu and C. R. Shelton. 2010. "Intrusion Detection using Continuous Time Bayesian Networks." *Journal of Artificial Intelligence Research*, 39, 745–774. doi:10.1613/jair.3050.

A Configuration and Data Dictionary

B Reproducibility Checklist for JAIR

Select the answers that apply to your research – one per item. Required by JAIR for all submissions (Gundersen et al. 2024).

All articles:

- (1) All claims investigated in this work are clearly stated. [yes]
- (2) Clear explanations are given how the work reported substantiates the claims. [yes]
- (3) Limitations or technical assumptions are stated clearly and explicitly. [yes]
- (4) Conceptual outlines and/or pseudo-code descriptions of the AI methods introduced in this work are provided, and important implementation details are discussed. [yes]
- (5) Motivation is provided for all design choices, including algorithms, implementation choices, parameters, data sets and experimental protocols beyond metrics. [partially]

Articles containing theoretical contributions

Does this paper make theoretical contributions? [no]

Articles reporting on computational experiments

Does this paper include computational experiments? [yes]

Table 15. Experimental configuration.

Item	Setting
Seeds	data 20260915; baselines 20260915; static graphs 20260927; temporal 20261003; ablations 20261004
Split	trajectory-level chronological 70/15/15
Positive class; threshold	UNAUTHORIZED = 1; 0.5
Raw categorical features	resource_type, action, clinical_purpose, tool_id, previous_action
Raw numeric features	resource_sensitivity, bulk_access_count_10m
Excluded fields	risk_score, authorization_label, attack_type, scenario_type, is_attack_event, purpose_consistency, patient_match, event and trajectory IDs
Context signals (Full)	patient scope (assigned = accessed patient); purpose-resource allow-list from authorized training events
Temporal model	frozen residual GAT encoder + causal forward GRU
Reference implementation	embedding 24; hidden 64; AdamW, lr 10^{-3} , weight decay 10^{-4} ; ≤ 30 epochs; patience 6; batch 256 events / 32 trajectories
Bootstrap	500 (per-model CI); 5,000 paired (ablation tests); unit = trajectory
Policy	Platt scaling; ALLOW < 0.45; DENY ≥ 0.92 ; least-privilege, patient- and purpose-scope overrides; constraints: unauthorized ALLOW $\leq 1\%$, authorized DENY $\leq 3\%$
Software	Python; pandas/NumPy; scikit-learn ≥ 1.5 ; XGBoost ≥ 2.0 ; PyTorch ≥ 2.3 ; SciPy ≥ 1.12 (exact versions and hardware to be reported with the code release)

- (1) All source code required for conducting experiments is included in an online appendix or will be made publicly available upon publication of the paper. The online appendix follows best practices for source code readability and documentation as well as for long-term accessibility. **[yes]**
- (2) The source code comes with a license that allows free usage for reproducibility purposes. **[yes]**
- (3) The source code comes with a license that allows free usage for research purposes in general. **[yes]**
- (4) Raw, unaggregated data from all experiments is included in an online appendix or will be made publicly available upon publication of the paper. The online appendix follows best practices for long-term accessibility. **[yes]**
- (5) The unaggregated data comes with a license that allows free usage for reproducibility purposes. **[yes]**
- (6) The unaggregated data comes with a license that allows free usage for research purposes in general. **[yes]**
- (7) If an algorithm depends on randomness, then the method used for generating random numbers and for setting seeds is described in a way sufficient to allow replication of results. **[partially]**
- (8) The execution environment for experiments, i.e., the computing infrastructure (hardware and software) used for running them, is described, including GPU/CPU makes and models; amount of memory (cache and RAM); make and version of operating system; names and versions of relevant software libraries and frameworks. **[partially]**
- (9) The evaluation metrics used in experiments are clearly explained and their choice is explicitly motivated. **[yes]**
- (10) The number of algorithm runs used to compute each result is reported. **[yes]**

Table 16. Fields of the security-event dataset.

Field	Definition	Model input
event_id, trajectory_id, task_id	identifiers	no
timestamp_utc	synthetic UTC time	ordering/splitting only
clinician_id, agent_id	synthetic clinician and agent	graph nodes
assigned/accessed_patient_id	task patient; patient accessed	nodes; patient scope
resource_id, resource_type, resource_name	FHIR-style resource	type, node identity
tool_id; action	tool READ/WRITE/DELETE/EXPORT	invoked; yes
clinical_purpose	authorized purpose	yes; purpose scope
resource_sensitivity	fixed per resource type (0–1)	yes
delegation_depth	always 0	no (constant)
previous_action;	prior action; 10-min access count	yes
bulk_access_count_10m		
purpose_consistency, patient_match, risk_score	generator helper/heuristic fields	no (leakage)
authorization_label	target	target only
attack_type, scenario_type, is_attack_event	scenario annotations	no; stratified reporting

- (11) Reported results have not been “cherry-picked” by silently ignoring unsuccessful or unsatisfactory experiments. **[yes]**
- (12) Analysis of results goes beyond single-dimensional summaries of performance (e.g., average, median) to include measures of variation, confidence, or other distributional information. **[yes]**
- (13) All (hyper-) parameter settings for the algorithms/methods used in experiments have been reported, along with the rationale or method for determining them. **[partially]**
- (14) The number and range of (hyper-) parameter settings explored prior to conducting final experiments have been indicated, along with the effort spent on (hyper-) parameter optimisation. **[no]**
- (15) Appropriately chosen statistical hypothesis tests are used to establish statistical significance in the presence of noise effects. **[partially]**

Articles using data sets

Does this work rely on one or more data sets (possibly obtained from a benchmark generator or similar software artifact)? **[yes]**

- (1) All newly introduced data sets are included in an online appendix or will be made publicly available upon publication of the paper. The online appendix follows best practices for long-term accessibility with a license that allows free usage for research purposes. **[yes]**
- (2) The newly introduced data set comes with a license that allows free usage for reproducibility purposes. **[yes]**
- (3) The newly introduced data set comes with a license that allows free usage for research purposes in general. **[yes]**

- (4) All data sets drawn from the literature or other public sources (potentially including authors' own previously published work) are accompanied by appropriate citations. **[yes]**
- (5) All data sets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. **[yes]**
- (6) All new data sets and data sets that are not publicly available are described in detail, including relevant statistics, the data collection process and annotation process if relevant. **[yes]**
- (7) All methods used for preprocessing, augmenting, batching or splitting data sets (e.g., in the context of hold-out or cross-validation) are described in detail. **[yes]**

Explanations on any of the answers above (optional)

Code and data licences (computational items 1–6, data items 1–3) assume release under an open licence (e.g., MIT and CC BY 4.0) upon publication. Seeds are reported per stage, but the random-number generators and some nondeterministic GPU operations are not fully specified (item 7). Library minimum versions are given; exact versions and hardware will accompany the code release (item 8). Final hyperparameters of the reference implementation are reported, but the search performed during development was not logged (items 13–14). Paired tests were applied to ablations but not to comparisons among separately trained baselines (item 15). The MedAgentBench task set is cited and publicly available; it served only as a design reference.

Received TBD; revised TBD; accepted TBD

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [MedTrustGraphStep9FinalPublicationBundle.zip](#)
- [ClinicalAgentInteractionGraphimage1.png](#)
- [ClinicalAIAgentDataAccessGraphimage1.png](#)