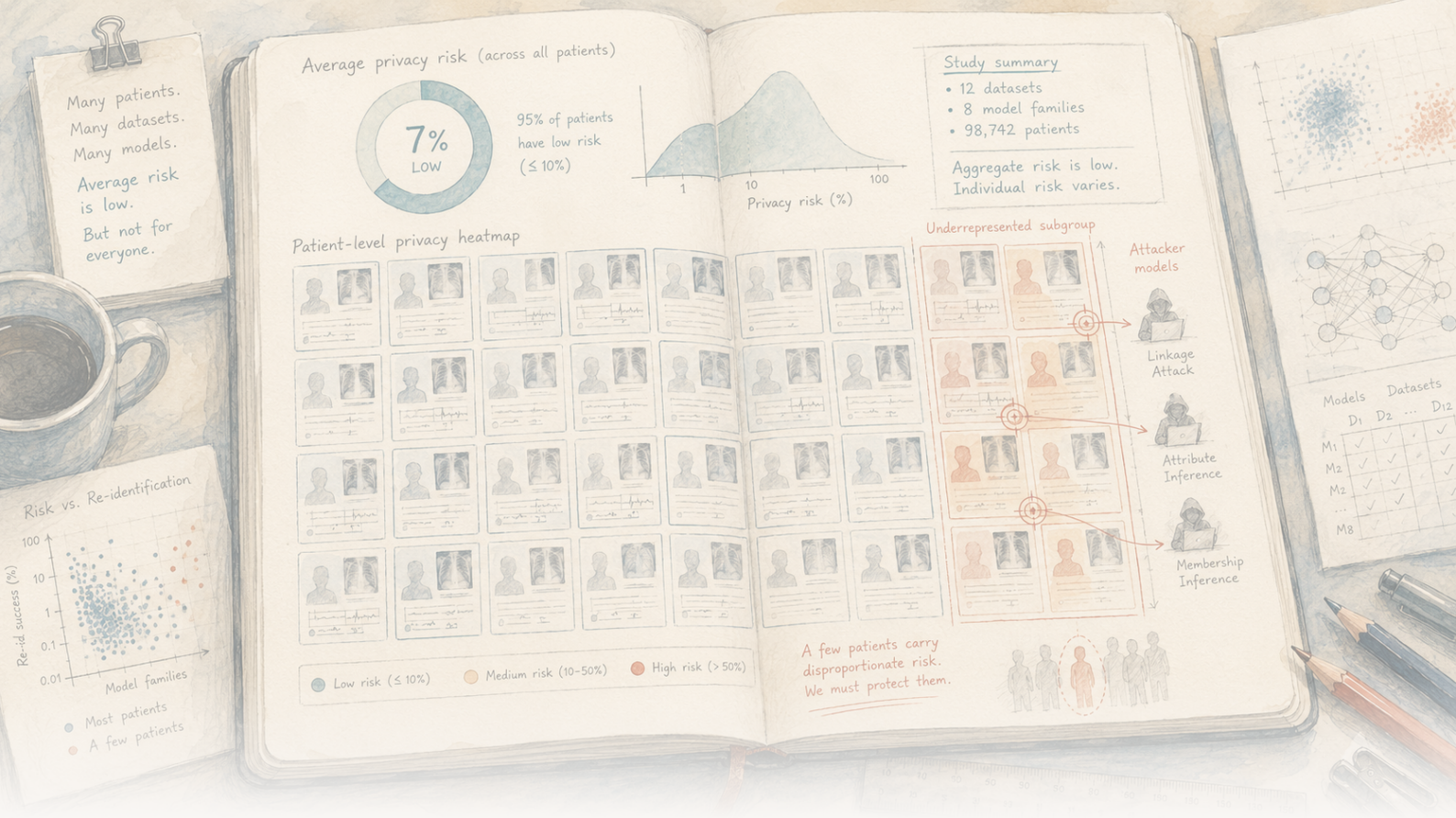


The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Medical AI average-এ private হয়েও নরিদষ্টি patient-কে expose করতে পারে — বশিষে করে underrepresented patient-কে

Seven medical dataset ও বহু model-এ per-patient membership-inference analysis দখোয় aggregate privacy metric severe individual risk hide করতে পার। কছিু patient-এর attack AUC ≥0.95 হলওে overall average chance-level, এবং extreme-risk tail-এ minority ethnicity, rare phenotype ও unusual image disproportionately বশেি। Model capacity বাড়লে tail risk বাড়ে। Study patient record content leak বা actual deployed attack frequency measure করে না; এটি দখোয় "private on average" ব্যক্তি-level guarantee নয়। Authors per-patient testing, access control ও differential privacy-এর মতে। mitigation চান।

Written by Lucio Vaglio · Cross-checked by Michele Renda · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Disparate privacy risks from medical AI*, Moritz Knolle, Georg Kaissis, Daniel Rückert and colleagues (Technical University of Munich; Imperial College London; Hasso Plattner Institute), Nature (2026) · DOI: 10.1038/s41586-026-10688-0

গোপনীয়তা নিশ্চয়তা গড়কে নয়, ব্যক্তি রোগীকে দেওয়া প্রতিশ্রুতি

চিকিৎসাবিষয়ক AI মডলেকে "গোপনীয়তা-সংরক্ষণকারী" বলা হলে প্রায়ই একটি সামগ্রিক গড় দখোনো হয়: প্রশিক্ষণ-তথ্যে থাকা রোগীদরে মধ্যে গড় সদস্যত্ব-অনুমতি সদ্ধিান্ত ঝুঁকি কম। শুনতে আশ্বস্তকর। এই গবষেণাপত্ররে অস্বস্তকির দকি: **গড় ভুল প্রশ্ন হতে পারে।**

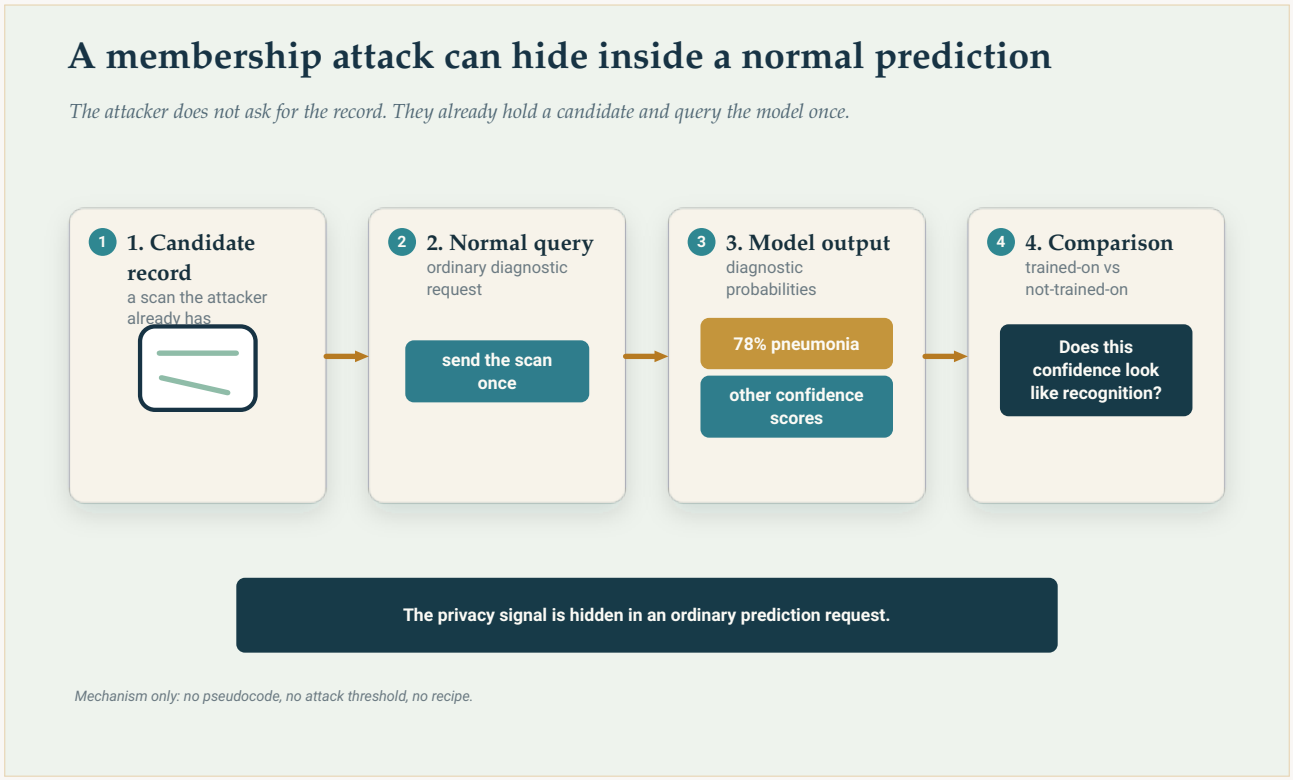
গোপনীয়তা একটি জনসংখ্যা গড় নয়; প্রতিটি ব্যক্তরি জন্য প্রতিশ্রুতি—প্রশিক্ষণ সটেে ছলি বলে যনে পরে তাকে শনাক্ত করা না যায়। গড় 999 রোগীকে সুরক্ষতি করে 1 জনকে প্রায় সম্পূর্ণভাবে উন্মুক্ত করওে নম্নি দখোতে পারে। গবষেকরো প্রথমবার প্রতিটি রোগীকে আলাদা করে ঝুঁকি পরমিাপ করে দখেছেনে: সামগ্রকি হসিাবে নরিাপদ বলে মনে হওয়া মডলে নর্দিষ্টি ব্যক্তরি সদস্যত্ব প্রায় নখিুঁতভাবে ফাঁস করতে পারে, এবং উন্মুক্ত ব্যক্তি অসমভাবে কম প্রতনিধিত্বিশীল গোষ্ঠীর সদস্য।

গবষেকরো কী করছেনে

Moritz Knolle, Daniel Rückert, Georg Kaissis ও সহকর্মীরা ভালোভাবে-পরচিতি **সদস্যত্ব অনুমতি সদ্ধিান্ত আক্রমণ**-এর প্রশ্ন বদলান। "ডটোসটে গড়ে আক্রমণ কতবার সফল?"-এর বদলে "**এই রোগী** কত উন্মুক্ত?"।

চিকিৎসাবিষয়ক ইমজেিং, ECG ও ইলকেট্রনকি স্বাস্থ্য নথি—ভন্নি তথ্য ধরনরে **7 প্রতিষ্ঠিতি চকিৎিসাবিষয়ক ডটোসটে**-এ, প্রতি ডটোসটেে বহু মডলে (গবষেণাপত্ররে বন্যিাসে 200 মডলে পর্যন্ত), রোগী-by-রোগী সদস্যত্ব ঝুঁকি অনুমান করা হয়। পরে রোগ অবস্থা, স্ব-প্রতবিদেতি জাতগিত পরচিয়, বমিা, লঙ্গি, ইমজেিং প্রোটোকল ইত্যাদি দল অনুযায়ী ভাগভত্তিকি বশি্লষেণ।

> **সদস্যত্ব অনুমতি সদ্ধিান্ত আক্রমণ আসলে কী**
>
> এখানে ফাঁস করা মানে মডলে "আপনার চকিৎিসাবিষয়ক নথি print করে দয়ে" নয়। ফাঁস করা হলো **সদস্যত্ব**—আপনার নথি প্রশিক্ষণ সটেে ছলি কি না।
>
> একটি রোগনির্ণয়মূলক মডলেকে স্ক্যান দলিে সে নডিমোনিয়া, cardiomegaly ইত্যাদরি সম্ভাবনা দয়ে। আক্রমণটি এই সাধারণ আউটপুটই ব্যবহার করে। প্রশিক্ষণে দখো উদাহরণে মডলেটি অদখো উদাহরণরে তুলনায় সামান্য বশেি আত্মবশি্বাসী হতে পারে—যমেন পরীক্ষার প্রশ্ন আগে দখো কোনো শকি্ষার্থী। আক্রমণকারীর কাছে সম্ভাব্য রোগীর নথি থাকে; সে মডলেকে স্বাভাবকি রোগনির্ণয়মূলক প্রশ্ন করে এবং আস্থার ধরনটি প্রশিক্ষণ-সদস্যরে মতো কি না দখে। রফোরন্সে বা shadow model ব্যবহার করে "আগে দখো" বনাম "অদখো" উদাহরণরে আস্থা-আচরণ শখো যায়।
>
> Request দখেতে স্বাভাবকি ক্লনিকি্যাল query-এর মতো; সংকতে পূর্বাভাস আস্থার ভতেরে। কনে সদস্যত্ব সংবদেনশীল? যদি মডলে বশিষে ক্যানসার চকিৎিসা নওেয়া রোগীর তথ্যে প্রশিক্ষণ দওেয়া হয়, কারো সদস্যত্ব অনুমান করা তার রোগ/চকিৎিসা সম্পর্কে সংবদেনশীল তথ্য reveal করতে পারে—এমনকি যদি নথি বষিয়বস্তু বরে না হয়। গবষেণা অনুমানগুলো—মডলে পূর্বাভাস প্রবশোধকিার, প্রার্থী নথি, আক্রমণকারী রফোরন্সে অবকাঠামো—খোলাখুলি দয়ে ঝুঁকি বিবিচেনামূলক চন্তিার জন্য।



Special গে াপনীয়তা API দরকার নইে: সাধারণ রে াগনর্িণয়মূলক পূর্বাভাসে প্রশক্ষিণ উদাহরণরে ওপর অস্বাভাবকি আস্থা সদস্যত্ব সংকতে তরৈি করতে পার।ে

Original Aurora diagram — The Clean Paper CC BY 4.0

তারা কী পয়েছেনে

গড় মান ঝুঁকরি লজে আড়াল করতে পার।ে সামগ্রকি মট্রেকিে বহু model-এর বরিুদ্ধে membership attack অধকিাংশ রে াগীর জন্য দবৈ অনুমানরে কাছাকাছি। কন্িতু ব্যক্তধিরে দখেলে কছিু রে াগীর ক্ষত্রেে attack AUC প্রায় নখুঁত—≥ 0.95; এখানে 0.5 মানে coin flip, 1.0 মানে নখুঁত বভিাজন। অর্থাৎ dataset-wide গড় নরীহ দখোলওে নর্িদষ্টি কয়কেজনরে ঝুঁকি খুব বশেি হতে পার।ে

গড় দবৈ সম্ভাবনার কাছে থাকতে পারে, অথচ ছোট subset-এর নথি অনকে বশেি উন্মুক্ত। গোপনীয়তা রোগী-স্তরে যাচাই না করলে লজে অদৃশ্য থাকে।

ঝুঁকি অসম, কম প্রতনিধিত্বিশীল গোষ্ঠীতে কনেদ্রীভূত করা। চরম-ঝুঁকি রোগী ধারাবাহকিভাবে সংখ্যালঘু জাতগিত পরচিয়, বরিলরোগ ফনেোটাইপ, অস্বাভাবকি ইমজেিং বশৈষ্ট্যিরে মতো কম প্রতনিধিত্বিশীল গোষ্ঠীতে বশেি। ইলকেট্রনকি-নথি ডটোসটেে কৃষ্ণ রোগী অধকিাংশ-ঝুঁকপ্রিবণ লজেে প্রত্যাশার তুলনায় **31% বশেি** দখো যায়; ম্যামোগ্রাফি ডটোসটেে malignancy-suspicious স্ক্যান ঝুঁকি লজেে **+1,179% আত্মীয় অতরিক্তি প্রতনিধিত্বি**। এগুলো ওই দলরে কত শতাংশ উন্মুক্ত—তা নয়; ছোট চরম-ঝুঁকি লজেে প্রত্যাশতি গণনার তুলনায় অতরিক্তি প্রতনিধিত্বি।

প্রক্রয়িাগত অন্তর্দৃষ্টি: মডলেরে কাছে সদৃশ উদাহরণ কম হলে অস্বাভাবকি নথি বশেি "stand বাইরে" করতে পারে; সদস্যত্ব আক্রমণ সইে মুখস্থকরণ/overconfidence সংকতে শনাক্ত করে। গোপনীয়তা ব্যর্থতা এলোমলেো নয়।

বড় মডলেে ঝুঁকি বাড়ে। মডলেরে সক্ষমতা বাড়ার সঙ্গে প্রায় নখিুঁতভাবে শনাক্তযোগ্য রোগীর সংখ্যা দ্রুত বাড়ে। একটি dermatology ডটোসটেে সবচয়েে ছোট মডলেে উচ্চ-ঝুঁকরি রোগী প্রায় ছলি না, কন্তিু সবচয়েে বড় মডলেে প্রায় **প্রতি 10 জনে 1 জন** উচ্চ-ঝুঁকরি লজেে পড়ে। চকিৎিসাবষিয়ক AI ক্রমে বড় হওয়ার প্রবণতায় এই নর্দিষ্টি গোপনীয়তা-ঝুঁককিে লখেকরো বাড়তে থাকা উদ্বগে হসিবেে দখেছেনে।

কনে "গড়ে নরিাপদ" ভুল পরীক্ষা

কম গড় ঝুঁকি মানইে সবার গোপনীয়তা নরিাপদ নয়। একটি গোপনীয়তা-নশ্চিয়তা শুধু "মধ্যম" রোগীর জন্য নয়, সবচয়েে বশেি উন্মুক্ত রোগীর ক্ষত্রেওে অর্থবহ হওয়া দরকার। 1,000 নথরি মধ্যে 999টি আলাদা করা না গলেওে যদি একটি নথকিে প্রায় নশ্চিতিভাবে শনাক্ত করা যায়, তবে "99.9% private" বলা সইে ব্যক্তরি জন্য অর্থহীন। আর যদি উচ্চ-ঝুঁকরি এই লজেে বরিল রোগ বা কম প্রতনিধিত্বিকারী গোষ্ঠীর রোগীরা বশেি থাকে, তবে গড়ভত্তিকি মট্রেকি শুধু অসম্পূর্ণ নয়—বষৈম্যও আড়াল করতে পারে।

গবষেণাপত্র উপস্থাপনভঙ্গি বদলায়: *মডলে গড়ে কত ব্যক্তগিত? থকেে সবচয়েে উন্মুক্ত রোগী কে, এবং তারা কোন দলরে?*। দ্বতিীয় প্রশ্ন

গোপনীয়তা নিশ্চয়তার সঙ্গে বেশি সামঞ্জস্যপূর্ণ।

এই গবেষণা কী প্রমাণ করে না

- চিকিৎসাবিষয়ক AI **ত্যাগ করা** করতে হবে—লেখকেরা প্রশমন চান, retreat নয়।
- প্রতিটি মোতায়নে করা চিকিৎসাবিষয়ক মডেল ফাঁস করছে বা breached হয়েছে—না; ঝুঁকি সক্ষমতা দেখানো হয়েছে।
- আপনার নথি ইতিমধ্যেই উন্মুক্ত—না; আক্রমণের নির্দিষ্ট প্রবেশাধিকার/নথি/রেফারেন্স অনুমান আছে।
- রোগী **বিষয়বস্তু** বের হয়ে আসে—এটি সদস্যত্ব ফাঁস করা, নথি dump নয়।
- বৈষম্য একটি একক সর্বজনীন সংখ্যা—না; দৃঢ় ফল হলো multi-ডেটাসেটে ধরন।

প্রমাণ কতটা শক্ত?

প্রায়োগিক পদ্ধতিগত ফল হিসেবে শক্তিশালী। সামগ্রিক গড়ভিত্তিক মেট্রিক ব্যক্তি ঝুঁকি understate করে, অবশিষ্ট ঝুঁকি কম প্রতিনিধিত্বশীল গোষ্ঠীতে কেন্দ্রীভূত করে এবং মডেল সক্ষমতার সঙ্গে worsens—ধরন **7 ভিন্ন চিকিৎসাবিষয়ক ডেটাসেটে** ও বহু মডেলে পুনরাবৃত্তি হয়েছে। Breadth এক-ডেটাসেটে কৃত্রিম প্রভাব ব্যাখ্যা দুর্বল করে।

দুটি সতর্কতা। প্রথমত এটি **ঝুঁকি প্রদর্শন**, বাস্তব জগতের ঘটনা কম্পাঙ্ক নয়; লেখকেরা নিজেদের আক্রমণ উল্লিখিত অনুমানগুলোর অধীনে চালিয়ে সক্ষম আক্রমণকারী কতটা সফল হতে পারে দেখেছেন। দ্বিতীয়ত চোখে পড়ার মতো প্রায় নখুঁত সংখ্যা ইচ্ছাকৃতভাবে সবচেয়ে খারাপ-উন্মুক্ত লেজে বর্ণনা করে; "অধিকাংশ রোগী উন্মুক্ত" নয়—বরং গড় লেজে কতটা আড়াল করে সেটিই বিন্দু।

সঠিক অবস্থান alarmism বা খারিজ নয়: মানক সামগ্রিক হিসাব গোপনীয়তা যাচাই সবচেয়ে খারাপ ক্ষেত্রগুলো বাদ পড়া করে, এবং সবচেয়ে খারাপ ক্ষেত্রগুলো সর্বনিম্ন-উপস্থাপিত রোগীতে বেশি পড়ে।

কেন এটি গুরুত্বপূর্ণ

প্রথম শিক্ষা: "গোপনীয়তা-সংরক্ষণকারী" বলার মানক বদলাতে হবে। প্রকাশের আগে **প্রতিরোগী ঝুঁকি** মূল্যায়ন করা, মডেল প্রবেশাধিকার নিয়ন্ত্রণ এবং কৌশল যেমন **পার্থক্যমূলক গোপনীয়তা** বিবেচনা করা দরকার। পার্থক্যমূলক গোপনীয়তা প্রশিক্ষণে ক্যালিব্রেটেড শব্দদূষণ/অ্যালগরিদমিক সীমাবদ্ধতা দিয়ে সদস্যত্ব সংকেত কমায়, কিন্তু গোপনীয়তা–utility সুবিধা-অসুবিধার বিনিময় বাস্তব; মুক্ত সুরক্ষা নয়। "গড় যাচাই করা" যথেষ্ট নয়।

দ্বিতীয় শিক্ষা হলো ন্যায্যতার প্রশ্নের সঙ্গে গোপনীয়তাকেও একসঙ্গে দেখা দরকার। চিকিৎসা-তথ্যে কম প্রতিনিধিত্বশীল গোষ্ঠী পূর্বাভাসের মানে ক্ষতিগ্রস্ত হতে পারে; একই গোষ্ঠী যদি গোপনীয়তার দিকি থেকেও সবচেয়ে কম সুরক্ষিত হয়, তবে দুটি বৈষম্যকে আলাদা বিভাগের সমস্যা হিসেবে দেখা যায় না।

সংক্ষপ্তি সারাংশ

Membership-inference attack নর্িধারণ করতে চায় কোনো নর্িদষ্টি রোগীর নথি model training data-তে ছলি কি না। সাতটি medical dataset—imaging, ECG ও EHR—এ ব্যক্তি-স্তররে ঝুঁকি মপেে গবষেণাটি দখোয়, সামগ্রকি গড় ঝুঁকি বভি্রান্তকির হতে পারে: dataset-এর গড় দবৈ অনুমানরে কাছাকাছি হলওে কছিু ব্যক্তরি attack AUC $\geq$**0.95**। উচ্চ-ঝুঁকরি এই লজে কম প্রতনিধিত্বিকারী গোষ্ঠী—সংখ্যালঘু জাতগিত পরচিয়, বরিল রোগ বা অস্বাভাবকি imaging—এ বশেি কনেদ্রীভূত; model বড় ও সক্ষম হলে সমস্যা আরও বাড়তে পারে। লখেকরো medical AI বন্ধ করতে বলনে না; তারা ব্যক্তি-স্তররে privacy audit, access control ও differential privacy ব্যবহাররে পক্ষে যুক্তি দনে। “গড়ে ব্যক্তগিত” হওয়া কোনো ব্যক্তরি গোপনীয়তার নশ্চিয়তা নয়।

সরাসরি যাচাই

পপোর যা দখোয়: Multi-ডটোসটে/মডলে গবষেণায় প্রতরিোগী সদস্যত্ব ঝুঁকি সামগ্রকি হসিাব গড়রে চয়েে কছিু ব্যক্ততিে অনকে বশেি; কম প্রতনিধিত্বিকারী দল অসমভাবে উন্মুক্ত; বড়তর মডলে ঝুঁকি বাড়ে।

যা সম্ভব, কন্তিু প্রমাণতি নয়: বাস্তব আক্রমণকারী মোতায়নে করা মডলে বর্তমানে একই আক্রমণ করছ—কম্পাঙ্ক গবষেণা পরমিাপ করনে।

যা দখোয় না: প্রতটিি চকিৎিসাবষিয়ক মডলে leaks, নর্িদষ্টি breach, নথি বষিয়বস্তু পর্যবক্ষেণকাল, চকিৎিসাবষিয়ক AI ত্যাগ করা উচতি।

প্রধান সীমাবদ্ধতা: লখেকরো-তরৈি আক্রমণ উল্লখিতি অনুমানগুলোর অধীন; সবচয়েে খারাপ-লজে মট্রেকি ইচ্ছাকৃতভাবে চরম ব্যক্তি মনোযোগ ধরে রাখা; সুনর্িদষ্টি দল বষৈম্য ডটোসটে-নর্িদষ্টি।

সাধারণ পাঠকরে আস্থা: সামগ্রকি গড়ভত্তিকি মট্রেকি ব্যক্তি ঝুঁকি বাদ পড়া করে এবং কম প্রতনিধিত্বিকারী রোগীর ওপর লজে কনেদ্রীভূত করতে পারে—উচ্চ। বাস্তব জগতরে আক্রমণ কম্পাঙ্ক—মাঝারি/অজানা।

উৎস

Knolle et al., Disparate privacy risks from medical AI, Nature (2026)
https://doi.org/10.1038/s41586-026-10688-0

Technical University of Munich — press release, ’Study reveals privacy risks in medical AI’ (2026)
https://www.tum.de/en/news-and-events/all-news/press-releases/details/study-reveals-privacy-risks-in

Medical Xpress (2026) — coverage of the study
https://medicalxpress.com/news/2026-06-patient-groups-vulnerable-privacy-medical.html