MEDECHO: PRIVACY-AWARE MULTI-AGENT COLLABORATION FOR CROSS-HOSPITAL DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical multi-agent systems combine diagnostic opinions, but their shared-case workflows do not directly fit cross-hospital diagnosis: raw cases remain at the originating hospital, while remote hospitals hold heterogeneous private knowledge. We identify two sequential failures in this setting: **restricting disclosure can lose diagnostic signal**, and **simple voting can miss a correct candidate already returned by collaborators**. We propose MEDECHO to address these stages. **Controlled Disclosure (CD)** builds, reviews, and audits a closed-schema collaboration packet, with local fallback when review fails. **Evidence Adoption (EA)** gathers source-tagged remote opinions and lets the originating hospital compare candidates against its complete case before issuing the final diagnosis. Extensive experiments on six datasets, three base LLM backbones, and three seeds show that MEDECHO improves dataset-equal accuracy over the strongest evaluated baseline by **15.94, 15.88, and 9.23 percentage points** on DeepSeek V4 Flash, Qwen 3.7 Flash, and GPT-5.6-Luna, respectively. These consistent backbone-level gains support the robustness and cross-backbone generalizability of MEDECHO within the evaluated diagnosis setting.

1 INTRODUCTION

Medical multi-agent systems (Medical MaS) bring specialized diagnostic agents together to reason over a case and combine their opinions (Kim et al., 2024; Tang et al., 2024; Chen et al., 2025b; Liu et al., 2025a). In common shared-case workflows, the collaborators can inspect the same case before their outputs are aggregated. Cross-hospital consultation has a different information structure: the raw case stays at the originating hospital, whereas remote hospitals contribute from separate, private knowledge resources. Multi-institutional privacy-preserving learning has long recognized the importance of keeping patient data at its source (Rieke et al., 2020; Sheller et al., 2020; Dayan et al., 2021), but case-level diagnosis also requires the hospitals to exchange enough information to consult on an individual patient. The originating hospital must therefore decide both what to disclose and what to do with the opinions it receives.

These decisions form two consecutive interfaces in the diagnostic workflow (Figure 1). At the first, a remote hospital cannot suggest a useful candidate if the released case description omits the clinical evidence it needs. At the second, discovering a correct candidate is insufficient if the final decision rule does not select it. This yields two concrete questions: **How can the originator release a bounded case description without discarding useful diagnostic signal? How can it adopt a useful remote candidate while retaining the full case and final diagnostic authority?** We focus on these two interfaces because they are necessary for a returned opinion to improve the final diagnosis, and because each has a measurable failure in the evaluated workflow. Other issues, including the quality of private knowledge and remote reasoning, can affect consultation but are not substitutes for either interface.

The matched voting comparison makes the first gap visible: replacing raw-context voting (RV) with packet voting (PV), while retaining the voting rule, is associated with an 8.22-percentage-point decrease in dataset-equal Accuracy (Acc), averaged over three base LLMs, six datasets, and three seeds. This **disclosure gap** reflects the evaluated communication-protocol change as a whole. The second gap appears within PV: the dataset-equal candidate-found rate is 64.64%, yet the final Top-1 Acc is only 41.25%, a 23.39-point **adoption gap**. Candidate availability and final selection are thus

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

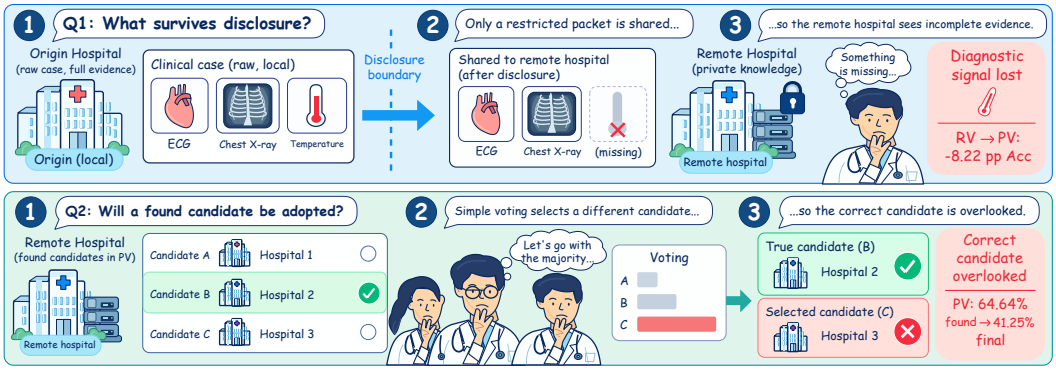

Figure 1: **The figure shows how omitted case signal and a missed correct candidate create two consecutive cross-hospital failures.** The clinical icons and voting example are schematic. The RV–PV Acc difference is 8.22 points; under PV, the candidate-found and final Top-1 rates are 64.64% and 41.25% (dataset-equal averages over six datasets, three backbones, and three seeds). CD and EA target the respective stages.

distinct. Together, these observations motivate a release contract that preserves useful case signal and a decision process that reconsiders returned candidates at the originator, rather than another voting rule alone.

We introduce MEDECHO to connect the two interfaces. **Controlled Disclosure (CD)** turns the raw case into a closed-schema collaboration packet, reviews it locally for allowed structure, plaintext leakage, and minimum clinical signal, records the release decision, and falls back to local diagnosis if review fails. **Evidence Adoption (EA)** obtains structured opinions from remote hospitals using their private retrieval collections, retains candidate provenance, and lets the originating hospital compare returned candidates with its complete case before making the final diagnosis. The coordinator only validates and forwards messages; it does not rank candidates or diagnose. Private retrieval supplies the heterogeneous knowledge setting, while CD and EA are the two method components that address the observed gaps.

Our contributions are summarized as follows.

❶ **Problem characterization.** We identify and quantify two sequential bottlenecks in the evaluated cross-hospital workflow: a disclosure-associated Acc decrease and a candidate-to-final-decision gap (Figure 1; Section 3.2).

❷ **Method.** We propose MEDECHO, coupling an auditable packet release and local fallback (CD) with source-tagged remote opinions and originator-side candidate comparison (EA; Section 3).

❸ **Evaluation.** Across six datasets, three base LLMs, and seeds 42/48/51, MEDECHO improves six-dataset-equal Acc over the strongest evaluated baseline by 15.94, 15.88, and 9.23 percentage points on DeepSeek V4 Flash, Qwen 3.7 Flash, and GPT-5.6-Luna, respectively; individual dataset–backbone results remain heterogeneous (Section 4).

2 RELATED WORK

2.1 MULTI-AGENT SYSTEM

Multi-agent research studies how language-model agents divide roles, exchange messages, and reach a final answer (Wang et al., 2024; Xi et al., 2025). Debate exposes and corrects reasoning errors (Du et al., 2024; Liang et al., 2024; Chan et al., 2024); role-playing and framework designs organize agents through explicit communication structures (Li et al., 2023; Hong et al., 2024; Wu et al., 2024; Qian et al., 2024; 2025; Chen et al., 2025a); and other systems add simulated environments (Park et al., 2023), verbal self-reflection and acting (Shinn et al., 2023; Huang et al., 2024a; Yao et al., 2023), or context partitioning and evidence exchange (Bo et al., 2024; Estornell et al., 2025), with quality measured on interactive benchmarks (Liu et al., 2024a; Chen et al., 2024a). Because agents act on

relayed content, collaboration also opens a trust boundary (Greshake et al., 2023; Debenedetti et al., 2024). MedECHO focuses on two stages that a cross-hospital exchange makes explicit: releasing case evidence and using the returned opinions.

2.2 Medical Multi-Agent System

Medical MaS applies these designs to clinical reasoning: MDAgents varies the collaboration strategy by task complexity (Kim et al., 2024), MedAgents assigns clinician roles (Tang et al., 2024), MAC aggregates opinions through conversation (Chen et al., 2025b), and MedChain sequences referral to treatment (Liu et al., 2025a). They build on medical language models and retrieval-augmented systems (Singhal et al., 2023; Thirunavukarasu et al., 2023; Tu et al., 2025; Hager et al., 2024; Xia et al., 2025) and are evaluated on examination-style questions (Jin et al., 2021), simulated environments (Mehandru et al., 2024; Gaber et al., 2025; Liu et al., 2025b; Goh et al., 2025; Schmidgall et al., 2026; Chen et al., 2026), and agent-driven science (Swanson et al., 2025). These studies motivate diagnostic collaboration; our setting additionally constrains what each hospital can see and who makes the final decision.

2.3 Privacy-Preserving Cross-Institutional Collaboration

Federated and decentralized learning keep training records local while coordinating models across institutions (Kairouz et al., 2021; Rieke et al., 2020; Sheller et al., 2020; Warnat-Herresthal et al., 2021; Dayan et al., 2021). This literature studies heterogeneous clients, privacy budgets, robustness, and communication security (Yang et al., 2024; Huang et al., 2023b; 2022; Chen et al., 2024b; Wan et al., 2024; Huang et al., 2023a; Yang et al., 2023; Liu et al., 2026; Huang et al., 2024b; Wei et al., 2025; Shen et al., 2025; Liu et al., 2024b; Shi et al., 2023; Zhang et al., 2025). Differential privacy and clinical de-identification address distinct exposure mechanisms (Dwork & Roth, 2014; Wei et al., 2020; Yu et al., 2022; Uzuner et al., 2007; Kaissis et al., 2020; Karargyris et al., 2023; Tayebi Arasteh et al., 2024; Bosma et al., 2025); attacks on federated learning and retrieval-augmented systems show why the released representation matters (Fowl et al., 2022; Qi et al., 2025). MedECHO operates at case-time consultation: CD reviews the message sent to remote hospitals, and EA specifies how their returned diagnoses enter the originating hospital's decision. These steps complement training-time privacy techniques rather than replace them.

3 Method

3.1 Problem Formulation

We consider cross-hospital diagnosis with one originating hospital h_o and multiple remote hospitals. The originator receives a raw case x and owns the final diagnosis; each remote hospital $j \in \mathcal{H} \setminus \{h_o\}$ has its own model, specialty coverage, and private retrieval collection R_j. The goal is to improve the primary diagnosis $\hat{y}$ with remote private knowledge while keeping the raw case at the originator and routing every remote contribution through a structured, auditable message. The target system is

$$\hat{y} = F_{\mathrm{EA}}(x, P(x), \{o_j\}_{j \neq h_o}), \tag{1}$$

where $P(x)$ is the reviewed packet released by CD and o_j is a structured remote opinion grounded in hospital j's local resources.

3.2 Motivation

To locate where cross-hospital collaboration loses performance, we compare two systems that differ only in what crosses the hospital boundary, holding datasets, seeds, and voting rule fixed. This yields the two observations in Figure 2.

Observation 1: replacing the raw case with a reviewed packet lowers the accuracy of the remote diagnoses. Raw-Context Voting (RV) votes over collaborators' primary diagnoses from the raw case; Packet Voting (PV) keeps the identical voting rule and changes only the transmitted content. With $A_b(M)$ the dataset-equal Acc of method M on backbone b, the difference between the two systems is the cost of the controlled communication protocol: 8.76 ± 1.15, 7.21 ± 1.08, and 8.68 ± 2.75 points

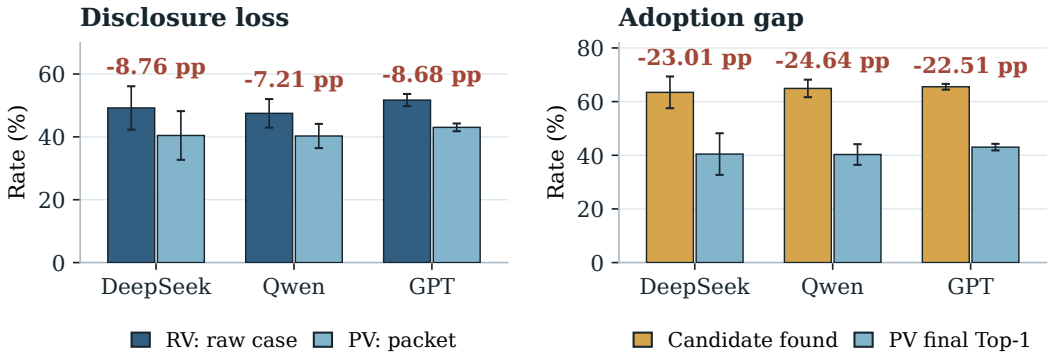

Figure 2: **The two panels measure accuracy lost at disclosure and the candidate-to-decision gap under packet voting.** Left: RV–PV dataset-equal Acc differences are 8.76/7.21/8.68 points for DeepSeek/Qwen/GPT. Right: candidate-found minus final Top-1 rates are 23.01/24.64/22.51 points. Bars show means and error bars show sample SD over three seeds; each seed average weights six datasets equally.

on DeepSeek, Qwen, and GPT (Figure 2, left). The boundary therefore removes clinical signal that collaborators were using.

Observation 2: a correct diagnosis does not become the final answer just because it has been found. Under packet voting the accepted diagnosis is present in the candidate set for 63.46/64.92/65.54% of cases but becomes the final primary diagnosis for only 40.45/40.28/43.03%, a gap of 23.01/24.64/22.51 points on DeepSeek/Qwen/GPT (Figure 2, right), whereas MEDECHO converts 90.11/88.26/88.26% of the candidates it finds into the final diagnosis, more than every baseline (Figure 10, Appendix D). Naming a candidate and committing to it are therefore different operations. The two gaps are complementary: the first concerns what crosses the boundary, the second what the originator does with what crossed. Together they set the requirement the paper addresses: keep the signal remote diagnosis needs, and let the originating hospital turn a found candidate into the final diagnosis.

3.3 OVERVIEW

Figure 3 shows the workflow: MEDECHO exposes three interfaces,

$$P(x) = \Phi_{\text{disc}}(x), \qquad \{o_j\}_{j\neq h_o} = \Phi_{\text{collab}}\big(P(x), \{R_j\}_{j\neq h_o}\big), \qquad \hat{y} = \Phi_{\text{adopt}}\big(x, \{o_j\}_{j\neq h_o}\big). \tag{2}$$

implemented by CD (Section 3.3.1) and EA (Section 3.3.2). The target system of Equation (1) therefore factorizes as

$$\hat{y} = F_{\text{EA}}(x, P(x), \{o_j\}_{j\neq h_o}) = \Phi_{\text{adopt}}\Big(x,\ \Phi_{\text{collab}}\big(\Phi_{\text{disc}}(x), \{R_j\}_{j\neq h_o}\big)\Big). \tag{3}$$

The raw case never leaves h_o: CD fixes what may cross the boundary, and EA decides which returned evidence replaces the local reading.

3.3.1 CD: CONTROLLED DISCLOSURE

CD targets the first failure mode: useful information must cross the hospital boundary without releasing the raw case. It projects x into a closed-schema COLLABORATIONPACKET containing patient context for risk stratification, clinical manifestations for symptoms and signs, objective-temporal evidence for examinations and course, and routing context for specialty selection. Hashed case and origin tags support auditability only; free-form narrative, direct identifiers, and undeclared fields are excluded, because released content from a local model or a retrieval index can be recovered verbatim (Fowl et al., 2022; Qi et al., 2025), and because utility-preserving safeguards remove identifiers, not clinical content (Kuo et al., 2025).

Let $p = \Pi_{\mathcal{K}}(x)$ denote projection onto the allowed field set $\mathcal{K}$. A local reviewer checks whether p follows the closed schema, contains no raw-case field or matching plaintext, and retains at least

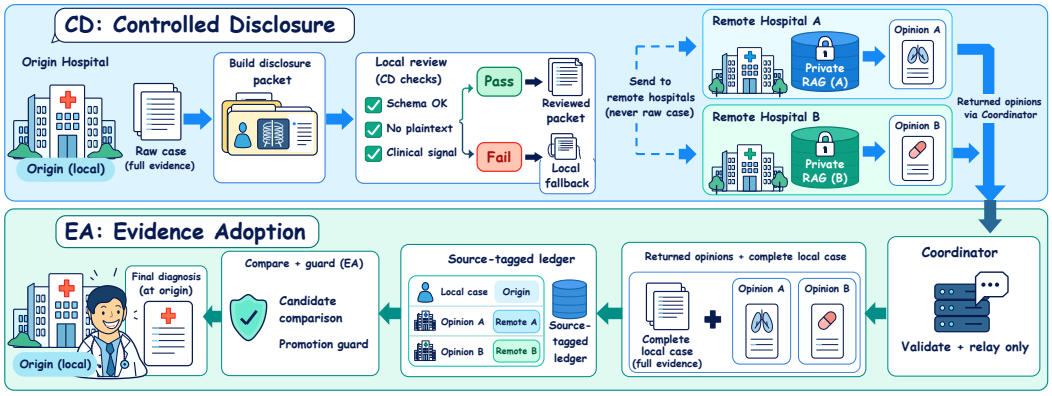

Figure 3: **The workflow shows how CD releases a locally reviewed packet and EA turns source-tagged remote opinions into an originator-side diagnosis.** Failed review invokes local fallback; remote hospitals use only the reviewed packet and their private retrieval collections. The coordinator validates and relays structured opinions without clinical ranking, while the raw case and final diagnostic authority remain at the originating hospital.

one controlled clinical concept or abnormal finding. Write the conjunction of these three checks as $\mathsf{Safe}(p, x)$. The released packet is

$$P(x) = \begin{cases} p, & \mathsf{Safe}(p, x), \\ \perp, & \text{otherwise.} \end{cases} \tag{4}$$

Each decision records the passed checks, hashed case reference, and fallback reason, and a rejected packet invokes local-only diagnosis. CD outputs an auditable disclosure decision; clinical scoring stays in the originator-side EA stage.

3.3.2 EA: Evidence Adoption

EA implements the two remaining interfaces and targets the second failure mode: remote hospitals first answer from their private resources (Φ_{collab}), and the originating hospital then re-evaluates the returned evidence (Φ_{adopt}).

Remote evidence construction. Each eligible hospital j queries its private R_j using only the reviewed packet and returns the structured opinion $o_j = (y_j, D_j, r_j, u_j) = f_j(P(x), R_j)$ of Equation 2, with a primary diagnosis, differential diagnoses, rationale, and confidence. The private retrieval collections instantiate the heterogeneous knowledge setting (Asai et al., 2024; Xia et al., 2025). All schema-valid opinions are relayed in stable hospital order; failed or invalid calls are logged and omitted, and the coordinator validates delivery and records failures.

Originator-side re-evaluation. Selected opinions return to h_o, where they are combined with the complete local case and organized into four label-free evidence channels, $\mathcal{E} = \{\text{existing Top-20}, \text{support RAG}, \text{HPO similarity}, \text{bounded text mention}\}$. Let $\mathcal{C}_e$ contain the candidates proposed by channel e. The candidate ledger and its two provenance statistics are

$$\mathcal{C} = \bigcup_{e \in \mathcal{E}} \mathcal{C}_e, \quad n(c) = \sum_{e \in \mathcal{E}} \mathbf{1}[c \in \mathcal{C}_e], \quad a(c) = \sum_{e \in \mathcal{E}} \mathbf{1}\left[c = \arg\max_{c' \in \mathcal{C}_e} q_e(c')\right], \tag{5}$$

where $n(c)$ counts evidence-channel presence and $a(c)$ counts channels for which c is the highest-scoring candidate; neither quantity counts hospitals. The incoming diagnosis y_{in} is the output of the preceding candidate-expansion and retrieval-aware selection stages. A frozen centroid ranker fitted on a separate development split scores the 60-dimensional, label-free evidence vector of each candidate as $s_\theta(c)$. With $c^* = \arg\max_{c \in \mathcal{C}} s_\theta(c)$ and $\Delta(c^*)$ denoting its margin over the second-ranked candidate, the originator applies

$$\hat{y} = \begin{cases} c^*, & c^* \neq y_{\text{in}} \ \wedge \ n(c^*) \geq 1 \ \wedge \ a(c^*) \geq m \ \wedge \ \Delta(c^*) \geq \tau, \\ y_{\text{in}}, & \text{otherwise.} \end{cases} \tag{6}$$

Here m is the minimum number of agreeing channels and τ is the promotion margin; candidate identities, case identities, and evaluation labels are excluded from s_θ. Separating candidate proposal from originator adoption lets EA keep a correct incoming diagnosis, promote a corroborated correction, or confirm agreement under conflicting evidence; candidate absence is handled upstream.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate six diagnosis datasets grouped by diagnostic complexity: **Simple** (S2D (Gretel AI, n.d.), K-S2D (Barman, n.d.), KDSD (Patil, n.d.)), **Difficult** (K-Sym (Godara, n.d.), Ext. (R. K., n.d.)), and **Rare** (MAC-302, a rare-disease subset built from a curated clinical case-report reference set (Caufield et al., 2018; Chen et al., 2025b)). Appendix C.1 gives dataset statistics, sampling, and per-seed test sizes.

Base LLMs. We evaluate DeepSeek V4 Flash (DeepSeek, 2026), Qwen 3.7 Flash (Qwen, 2026), and GPT-5.6-Luna (OpenAI, 2026). Each run uses one base LLM at every hospital node, the same graph and disclosure protocol, and temperature 0.

Compared Methods. We compare three method families: local diagnosis (**LR**, retrieval-augmented diagnosis at the originating hospital (Lewis et al., 2020)), shared-context collaboration (**RV**, an independent-opinion vote over the raw case (Du et al., 2024), plus **MAC** (Chen et al., 2025b) and **MedChain** (Liu et al., 2025a)), and controlled-disclosure collaboration (**PV**, the RV rule applied to reviewed packets, and **MedECHO**). Appendix C gives the full contracts.

Metrics. We use Accuracy (Acc) and Error Correction Rate (ECR) as the main metrics. For N matched cases, let Y_i be the accepted reference-label set, let $\hat{y}_i^M$ be method M's final primary diagnosis after canonical label normalization, and define per-case correctness $\zeta_i^M = \mathbb{1}[\hat{y}_i^M \in Y_i]$. We report

$$\mathrm{Acc}(M) = \frac{1}{N}\sum_{i=1}^{N} \zeta_i^M, \tag{7}$$

$$\mathrm{ECR}(M) = \frac{\sum_{i=1}^{N}(1-\zeta_i^{\mathrm{LR}})\zeta_i^M}{\sum_{i=1}^{N}(1-\zeta_i^{\mathrm{LR}})}, \tag{8}$$

where ζ_i^{LR} is the correctness of the matched LR reference and the ECR denominator is its number of errors, so LR's ECR is zero by construction. Every method in a cell shares the matched case set and LR reference, so the comparison never mixes cohorts. Appendix B.3 gives the full decomposition.

4.2 MAIN RESULTS

MedECHO achieves the strongest dataset-equal Acc and ECR averages on all three base LLMs. Figure 4 reports every cell and Appendix D the exact values.

Each gain below is measured against the strongest baseline of that group: MAC on Simple for DeepSeek and Qwen, LR on Simple for GPT, and RV on Difficult+Rare for all three. On **Simple**, MedECHO reaches 88.14/86.33/87.37% Acc with DeepSeek/Qwen/GPT, improving over the strongest baseline group average by 15.23/19.47/14.73 points. On **Difficult+Rare**, it reaches 42.16/40.41/34.51%, with gains of 14.27/11.95/3.64 points. MAC-302 is the smallest cohort (16/17/16 cases over the three seeds), so its intervals are widest. Averaged over the six datasets, MedECHO reaches 65.15/63.37/60.94% Acc, 15.94/15.88/9.23 points above the strongest baseline. ECR rises from 9.34/7.05/8.64 under RV to 53.15/49.64/46.22 under MedECHO.

4.3 AGENT COMPOSITION

Table 1 fixes GPT-5.6-Luna at the originator and varies the two remote models: homogeneous (Luna/Luna), same-family (Terra/Sol), and cross-family (DeepSeek/Qwen). This study uses seed 42 on S2D, K-Sym, and MAC-302. MedECHO has the highest average ECR in the homogeneous and

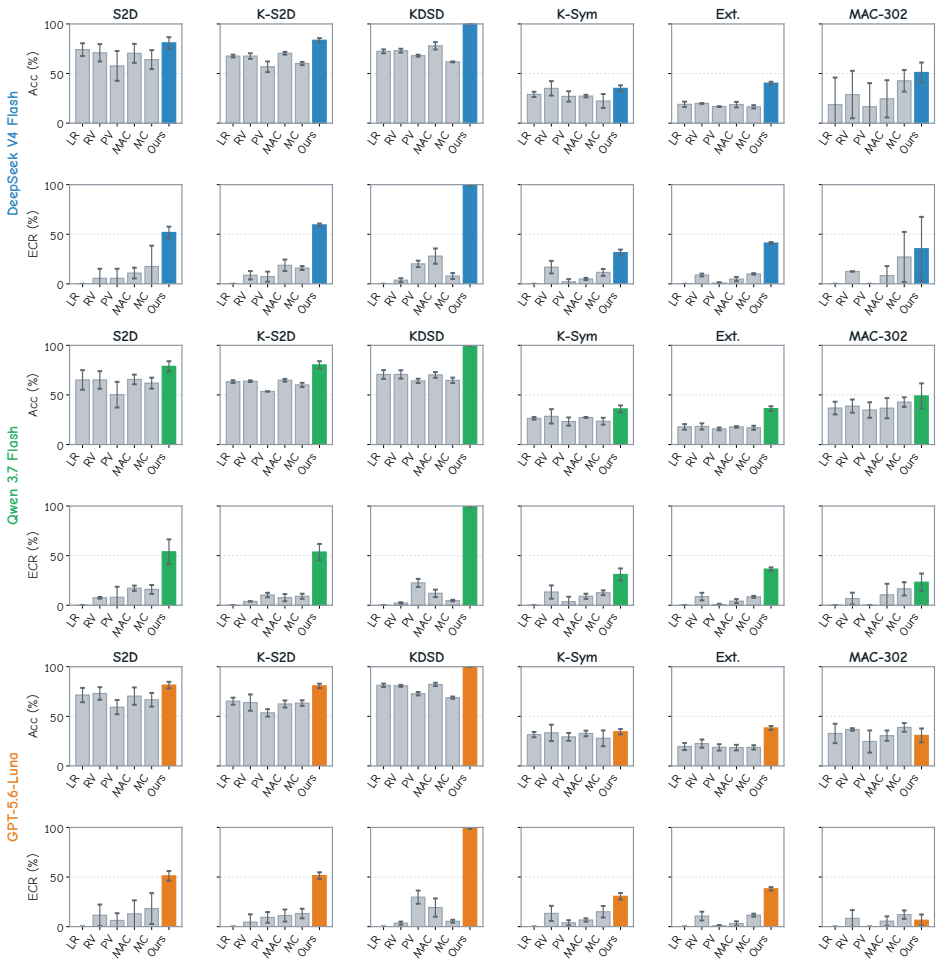

Figure 4: **The plot compares dataset-equal Acc and ECR across methods and all three base LLMs.** Bars show three-seed means with sample SD; MEDECHO has the highest backbone-level averages in both metrics.

cross-family settings and the highest average Acc in the cross-family setting. RV leads average Acc in the other two settings, while MAC leads same-family ECR; MedECHO's MAC-302 ECR remains zero in all three.

4.4 Ablation Experiments

All analyses in this subsection reuse sealed predictions and make no additional model calls.

CD: privacy versus retained signal. On six canaries with 13 declared clinical signals (Figure 5), CD retains all signals with zero measured exposure. Raw sharing and masking expose 100% and 50%, respectively; minimal keywords reach zero exposure but retain only 46.15% of signals.

Decision-rule hyperparameters. DeepSeek replays on S2D, K-Sym, and Ext. show unchanged Acc across the tested $\tau \in \{0, .05, .10\}$ (Figure 6); the deployed margin is $\tau = .15$. Reducing agreeing channels from $m = 2$ to 1 lowers K-Sym/Ext. Acc by 23.48/23.65 points; increasing to 3 adds only 0.87/1.35 points.

EA ablation. Table 2 builds the decision stack step by step on the six-dataset cohort.

Table 1: **Changing remote models tests agent composition with a fixed GPT-5.6-Luna originator.**; Avg weights the three datasets equally. Triples list originator / remote A / remote B.

Methods	S2D		K-Sym		MAC-302		Avg	
	Acc	ECR	Acc	ECR	Acc	ECR	Acc	ECR
Luna / Luna / Luna								
LR	77.78	0.00	39.13	0.00	31.25	0.00	49.39	0.00
RV	79.37	21.43	41.74	14.29	37.50	9.09	52.87	14.94
PV	65.08	14.29	35.65	4.29	25.00	0.00	41.91	6.19
MAC	79.37	35.71	34.78	4.29	25.00	0.00	46.38	13.33
MedChain	71.43	35.71	34.78	17.14	37.50	9.09	47.90	20.65
MedECHO (Ours)	**90.48**	**57.14**	**53.04**	**31.43**	**43.75**	**18.18**	**62.42**	**35.58**
Luna / Terra / Sol								
LR	77.78	0.00	39.13	0.00	31.25	0.00	49.39	0.00
RV	87.30	50.00	43.48	21.43	37.50	9.09	56.09	26.84
PV	73.02	28.57	34.78	5.71	25.00	0.00	44.27	11.43
MAC	88.89	64.29	34.78	5.71	43.75	18.18	55.81	29.39
MedChain	74.60	28.57	29.57	11.43	31.25	0.00	45.14	13.33
MedECHO (Ours)	**95.24**	**71.43**	**52.17**	**30.00**	**50.00**	**27.27**	**65.80**	**42.90**
Luna / DeepSeek / Qwen								
LR	77.78	0.00	39.13	0.00	31.25	0.00	49.39	0.00
RV	74.60	7.14	34.78	14.29	43.75	18.18	51.05	13.20
PV	61.90	14.29	37.39	8.57	31.25	0.00	43.52	7.62
MAC	77.78	42.86	36.52	7.14	31.25	9.09	48.52	19.70
MedChain	76.19	50.00	30.43	12.86	37.50	9.09	48.04	23.98
MedECHO (Ours)	**93.65**	**57.14**	**53.04**	**30.00**	**50.00**	**27.27**	**65.56**	**38.14**

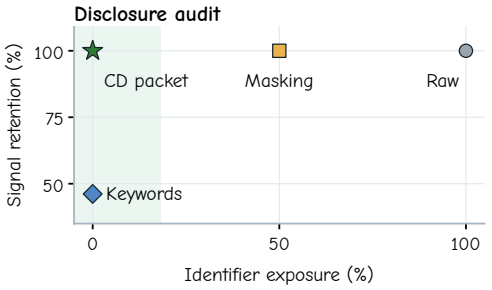

Figure 5: **Release strategies trade identifier exposure against signal retention.** Six canaries, 13 DeepSeek on S2D, K-Sym, and Ext.; dashed line: declared signals; CD retains all 13 with zero measured exposure.

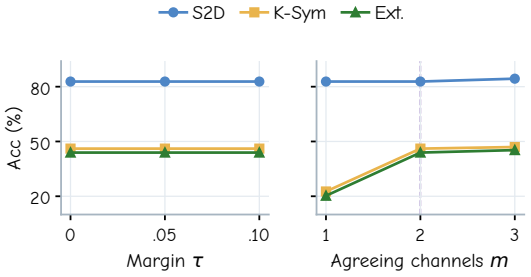

Figure 6: **One-factor sweeps test decision sensitivity.** played $m = 2$.

Adding the candidate ledger contributes the largest Acc gain: **16.37/15.94 points** on DeepSeek/Qwen. Retrieval-aware selection adds 1.32/0.31 points and the guard adds 0.60/0.10. Leave-one-out replays lose 14.30/12.56 points without candidate expansion and 12.14/11.68 without Support RAG; removing either lexical channel changes Acc by at most 1.11 points (Appendix D).

Table 2: **The table adds EA components cumulatively and reports six-dataset-equal Acc (%).** These replays use the pre-filter ablation cohort.

Cumulative configuration	DeepSeek	Qwen
Remote opinions only	48.44	50.55
+ candidate ledger	64.81	66.49
+ retrieval-aware selection	66.13	66.80
+ guard (full EA)	66.73	66.90

Efficiency against multi-agent baselines.
Figure 7 compares accuracy against token workload. MedECHO reaches 51.98% Acc with 502,910 normalized Tokens and is non-dominated with LR and RV. MAC reaches 55.81% Acc with 1.36M Tokens, and MedChain reaches 45.14% with 1.55M Tokens. These seed-42 same-family results average S2D, K-Sym, and MAC-302. The online stage ends once opinions are sealed; candidate expansion, selection, and scoring run locally (Appendix D).

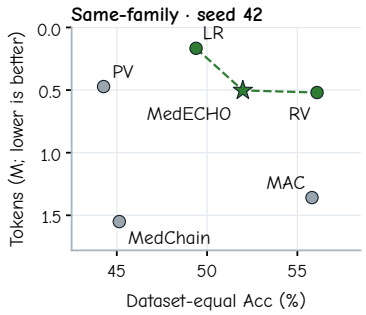

Figure 7: **Same-family composition trades Acc against token workload.** Seed 42; S2D, K-Sym, MAC-302. LR tokens are a lower bound; dashed line: non-dominated methods.

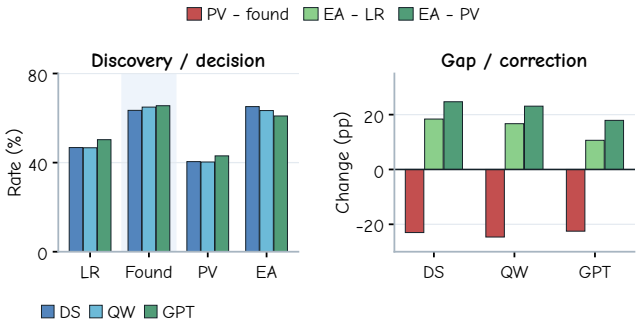

Figure 8: **The panels separate candidate discovery from final diagnosis.** Six dataset-equal means, three seeds. Found: PV candidate availability; EA: MedECHO final Acc. DS/QW: DeepSeek/Qwen.

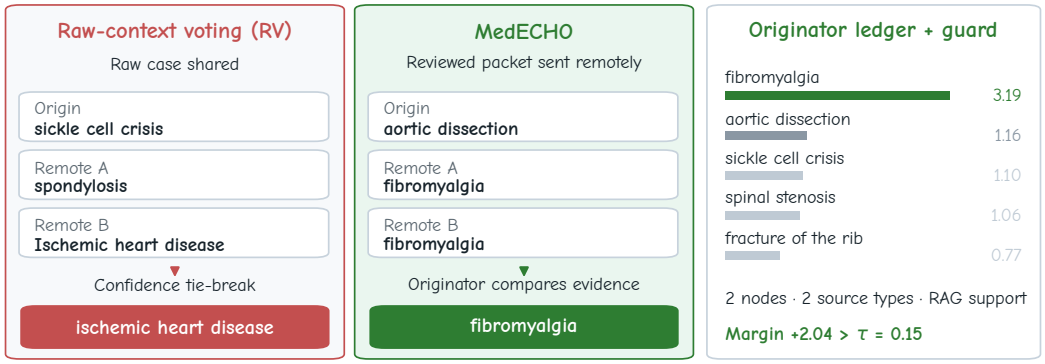

Figure 9: **The case trace compares RV voting with originator-side evidence adoption.** Ext., DeepSeek, seed 51; reference diagnosis: fibromyalgia. The ledger shows five candidate scores and the promotion guard.

4.5 In-Depth Analysis

Case-level decision trace. Figure 9 follows one case through both decision rules. RV receives three different diagnoses and resolves the tie on confidence, producing an incorrect answer. With a reviewed packet, both remote hospitals instead propose fibromyalgia. EA records their source-tagged support; the candidate's $+2.04$ margin exceeds $\tau = 0.15$, and the originator promotes it over its local reading of aortic dissection.

Why provenance matters. Voting treats opinions as interchangeable and cannot compare them with the full case retained at the originator. EA retains candidate sources for that comparison and bounds promotion by corroboration. In the tested sensitivity sweep, reducing the required evidence-channel count from two to one lowers Acc by more than 20 points on K-Sym and Ext. (Figure 6).

5 Conclusion

MedECHO addresses two consecutive failures in cross-hospital diagnosis: CD reviews what leaves the originating hospital, and EA turns source-tagged remote opinions into a final decision at that hospital. Across six datasets, three base LLMs, and three seeds, dataset-equal Acc exceeds the strongest baseline by 15.94/15.88/9.23 points. The candidate-ledger ablation identifies the largest contribution within EA; a six-canary disclosure audit finds all 13 declared signals retained with no identifier exposure. These results support a collaboration contract in which remote hospitals contribute evidence while the originator retains the raw case and diagnostic authority.

AI USE STATEMENT

Generative AI assisted methodology feedback, code implementation, experiment orchestration, data reformatting, result analysis, figure generation, translation, and manuscript drafting. It was not used to generate evaluation labels. The authors remain responsible for verifying the code, tests, frozen artifacts, numerical claims, citations, figures, and final text.

REPRODUCIBILITY STATEMENT

All reported results use seeds 42/48/51 and exclude every source case that appears in the development split used to fit the decision policy. Exporters verify completion, matched case keys, schemas, and hashes; offline replay and paper export make zero API calls and load no model.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Niyarr Barman. Symptom2disease. Kaggle, n.d. URL https://www.kaggle.com/datasets/niyarrbarman/symptom2disease. Accessed: 2026-08-03.

Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. Reflective multi-agent collaboration based on large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4397. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fa54b0edce5eef0bb07654e8ee800cb4-Abstract-Conference.html.

Joeran S. Bosma, Koen Dercksen, Luc Builtjes, Romain Andre, Christian Roest, Stefan J. Fransen, Constant R. Noordman, Mar Navarro-Padilla, Judith Lefkes, Natalia Alves, Max J. J. de Grauw, Leander van Eekelen, Joey M. A. Spronck, Megan Schuurmans, Bram de Wilde, Ward Hendrix, Witali Aswolinskiy, Anindo Saha, Jasper J. Twilt, Daan Geijs, Jeroen Veltman, Derya Yakar, Maarten de Rooij, Francesco Ciompi, Alessa Hering, Jeroen Geerdink, Henkjan Huisman, and DRAGON Consortium. The DRAGON benchmark for clinical NLP. *npj Digital Medicine*, 8:289, 2025. doi: 10.1038/s41746-025-01626-x. URL https://www.nature.com/articles/s41746-025-01626-x.

J. Harry Caufield, Yijiang Zhou, Anders O. Garlid, Shaun P. Setty, David A. Liem, Quan Cao, Jessica M. Lee, Sanjana Murali, Sarah Spendlove, Wei Wang, Li Zhang, Yizhou Sun, Alex Bui, Henning Hermjakob, Karol E. Watson, and Peipei Ping. A reference set of curated biomedical data and metadata from clinical case reports. *Scientific Data*, 5:180258, 2018. doi: 10.1038/sdata.2018.258.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Eric Chen, Sam Postelnik, Kameron Black, Yixing Jiang, and Jonathan H. Chen. MedAgentBench v2: Improving medical LLM agent design. In *Pacific Symposium on Biocomputing (PSB)*, pp. 354–371, 2026. doi: 10.1142/9789819824755_0025.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations (ICLR)*, 2024a. URL https://openreview.net/forum?id=EHg5GDnyq1.

Weize Chen, Ziming You, Ran Li, Yitong Guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://openreview.net/forum?id=o1Et3MogPw.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, Qicheng Lao, Weili Fu, Kang Li, and Jian Li. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Medicine*, 8:159, 2025b. doi: 10.1038/s41746-025-01550-0. URL https://www.nature.com/articles/s41746-025-01550-0.

Yuhang Chen, Wenke Huang, and Mang Ye. Fair federated learning under domain skew with local consistency and domain diversity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12077–12086, 2024b. doi: 10.1109/CVPR52733.2024.01148.

Ittai Dayan, Holger R. Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z. Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J. Wood, Chien-Sung Tsai, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, 27(10): 1735–1743, 2021. doi: 10.1038/s41591-021-01506-3.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 82895–82920, 2024. doi: 10.52202/079017-2636.

DeepSeek. DeepSeek V4 Flash, 2026. URL https://api-docs.deepseek.com/news/news260424/. Accessed: 2026-08-03.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–487, 2014. doi: 10.1561/0400000042.

Andrew Estornell, Jean-Francois Ton, Yuanshun Yao, and Yang Liu. ACC-Collab: An actor-critic approach to multi-agent LLM collaboration. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=nfKfAzkiez.

Liam H Fowl, Jonas Geiping, Wojciech Czaja, Micah Goldblum, and Tom Goldstein. Robbing the fed: Directly obtaining private data in federated learning with modified models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=fwzUgo0FM9v.

Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke, and Altuna Akalin. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8:263, 2025. doi: 10.1038/s41746-025-01684-1. URL https://www.nature.com/articles/s41746-025-01684-1.

Abhishek Godara. Symptoms to diseases. Kaggle, n.d. URL https://www.kaggle.com/datasets/abhishekgodara/symptoms-to-diseases. Accessed: 2026-08-03.

Ethan Goh, Robert J. Gallo, Eric Strong, Yingjie Weng, Hannah Kerman, Jason A. Freed, Josephine A. Cool, Zahir Kanjee, Kathleen P. Lane, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Jason Hom, Jonathan H. Chen, and Adam Rodman. GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial. *Nature Medicine*, 31:1233–1238, 2025. doi: 10.1038/s41591-024-03456-y. URL https://www.nature.com/articles/s41591-024-03456-y.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, pp. 79–90, 2023. doi: 10.1145/3605764.3623985.

Gretel AI. Symptom to diagnosis. Hugging Face Datasets, n.d. URL https://huggingface.co/datasets/gretelai/symptom_to_diagnosis. Accessed: 2026-08-03.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622, 2024. doi: 10.1038/s41591-024-03097-1.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys*, 54(4):1–37, 2021. doi: 10.1145/3447772.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024a. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10133–10143, 2022. doi: 10.1109/CVPR52688.2022.00990.

Wenke Huang, Guancheng Wan, Mang Ye, and Bo Du. Federated graph semantic and structural learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3830–3838, 2023a. doi: 10.24963/ijcai.2023/426.

Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16312–16322, 2023b. doi: 10.1109/CVPR52729.2023.01565.

Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, Bo Du, and Qiang Yang. Federated learning for generalization, robustness, fairness: A survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9387–9406, 2024b. doi: 10.1109/TPAMI.2024.3418862.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421.

Janet Joy and Andrew I. Su. Federated knowledge retrieval elevates large language model performance on biomedical benchmarks. *GigaScience*, 15, 2026. doi: 10.1093/gigascience/giag007.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14 (1–2):1–210, 2021. doi: 10.1561/2200000083.

Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2:305–311, 2020. doi: 10.1038/s42256-020-0186-1. URL https://www.nature.com/articles/s42256-020-0186-1.

Alexandros Karargyris, Renato Umeton, Micah J. Sheller, Alejandro Aristizabal, Johnu George, Anna Wuest, Sarthak Pati, Hasan Kassem, Maximilian Zenk, Ujjwal Baid, Prakash Narayana Moorthy, Alexander Chowdhury, Junyi Guo, Sahil Nalawade, Jacob Rosenthal, David Kanter, Maria Xenochristou, Daniel J. Beutel, Verena Chung, Timothy Bergquist, James Eddy, Abubakar Abid, Lewis Tunstall, Omar Sanseviero, Dimitrios Dimitriadis, Yiming Qian, Xinxing Xu, Yong Liu,

Rick Siow Mong Goh, Srini Bala, Victor Bittorf, Sreekar Reddy Puchala, Biagio Ricciuti, Soujanya Samineni, Eshna Sengupta, Akshay Chaudhari, Cody Coleman, Bala Desinghu, Gregory Diamos, Debo Dutta, Diane Feddema, Grigori Fursin, Xinyuan Huang, Satyananda Kashyap, Nicholas Lane, Indranil Mallick, FeTS Consortium, BraTS-2020 Consortium, AI4SafeChole Consortium, Pietro Mascagni, Virendra Mehta, Cassiano Ferro Moraes, Vivek Natarajan, Nikola Nikolov, Nicolas Padoy, Gennady Pekhimenko, Vijay Janapa Reddi, G. Anthony Reina, Pablo Ribalta, Abhishek Singh, Jayaraman J. Thiagarajan, Jacob Albrecht, Thomas Wolf, Geralyn Miller, Huazhu Fu, Prashant Shah, Daguang Xu, Poonam Yadav, David Talby, Mark M. Awad, Jeremy P. Howard, Michael Rosenthal, Luigi Marchionni, Massimo Loda, Jason M. Johnson, Spyridon Bakas, and Peter Mattson. Federated benchmarking of medical artificial intelligence with MedPerf. *Nature Machine Intelligence*, 5:799–810, 2023. doi: 10.1038/s42256-023-00652-2. URL https://www.nature.com/articles/s42256-023-00652-2.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. ProPILE: Probing privacy leakage in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20750–20762, 2023. doi: 10.52202/075280-0911.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2522. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/90d1fc07f46e31387978b88e7e057a31-Abstract-Conference.html.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M. Brower, et al. The human phenotype ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217, 2021. doi: 10.1093/nar/gkaa1043.

Martin Kuo, Jingyang Zhang, Jianyi Zhang, Minxue Tang, Louis DiValentin, Aolin Ding, Jingwei Sun, William Chen, Amin Hass, Tianlong Chen, Yiran Chen, and Hai Li. Proactive privacy amnesia for large language models: Safeguarding PII with negligible impact on model utility. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=io8uRPYktn.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2264. URL https://proceedings.neurips.cc/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html.

Yapeng Li, Yong Luo, and Bo Du. UVaT: Uncertainty incorporated view-aware transformer for robust multi-view classification. *IEEE Transactions on Image Processing*, 33:5129–5143, 2024a. doi: 10.1109/TIP.2024.3451931.

Yapeng Li, Yong Luo, Zengmao Wang, and Bo Du. Improving generalized zero-shot learning by exploring the diverse semantics from external class names. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23344–23353, 2024b. doi: 10.1109/CVPR52733.2024.02203.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 17889–17904, 2024. doi: 10.18653/v1/2024.emnlp-main.992.

Chang Liu, Yapeng Li, Zengmao Wang, Dewen Soh, and Bo Du. MAPFL: Modular adaptive personalised federated learning from visual sentiment to physical activity. *CAAI Transactions on Intelligence Technology*, 2026. doi: 10.1049/cit2.70120.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Haoliang Li, Linlin Shen, Michael R. Lyu, and Wenting Chen. MedChain: Bridging the gap between LLM agents and clinical practice with interactive sequence. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6bc02900c74234aa5a0c1a7148ed933f-Abstract-Datasets_and_Benchmarks_Track.html.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*, 2024a. URL https://openreview.net/forum?id=zAdUB0aCTQ.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, Tianpei Hong, Jin Yang, Tianrun Gao, Jiangjiang Zhang, Xiaohu Li, Jing Zhang, Ye Sang, Zhao Yang, Kanmin Xue, Song Wu, Ping Zhang, Jian Yang, Chunli Song, and Guangyu Wang. A generalist medical language model for disease diagnosis assistance. *Nature Medicine*, 31:932–942, 2025b. doi: 10.1038/s41591-024-03416-6. URL https://www.nature.com/articles/s41591-024-03416-6.

Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-Qin Zhang, and Qiang Yang. Vertical federated learning: Concepts, advances, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3615–3634, 2024b. doi: 10.1109/TKDE.2024.3352628.

Nikita Mehandru, Brenda Y. Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J. Butte, and Ahmed Alaa. Evaluating large language models as agents in the clinic. *npj Digital Medicine*, 7:84, 2024. doi: 10.1038/s41746-024-01083-y. URL https://www.nature.com/articles/s41746-024-01083-y.

OpenAI. GPT-5.6-Luna model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna. Accessed: 2026-08-03.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763.

Pranay Patil. Disease symptom prediction. Kaggle, n.d. URL https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset. Accessed: 2026-08-03.

Zhenting Qi, Hanlin Zhang, Eric P Xing, Sham Kakade, and Hima Lakkaraju. Follow my instruction and spill the beans: Scalable data extraction from retrieval-augmented generation systems. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=Y4aWwRh25b.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.810. URL https://aclanthology.org/2024.acl-long.810/.

Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large language model-based multi-agent collaboration. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=K3n5jPkrU6.

Qwen. Model changelog, 2026. URL https://platform.qianwenai.com/docs/changelog/models. Accessed: 2026-08-03.

Qwen Team. Qwen3-Embedding-4B. Hugging Face model card, 2025. URL https://huggingface.co/Qwen/Qwen3-Embedding-4B. Accessed: 2026-09-19.

Dhivyesh R. K. Disease-symptom dataset. Kaggle, n.d. URL https://www.kaggle.com/datasets/dhivyeshrk/diseases-and-symptoms-dataset. Accessed: 2026-08-03.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *npj Digital Medicine*, 3:119, 2020. doi: 10.1038/s41746-020-00323-1.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*, 9(1), 2026. doi: 10.1038/s41746-026-02674-7.

Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, and Spyridon Bakas. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1), 2020. doi: 10.1038/s41598-020-69250-1.

Wei Shen, Wenke Huang, Guancheng Wan, and Mang Ye. Label-free backdoor attacks in vertical federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20389–20397, 2025. doi: 10.1609/aaai.v39i19.34246.

Yujun Shi, Jian Liang, Wenqing Zhang, Vincent Tan, and Song Bai. Towards understanding and mitigating dimensional collapse in heterogeneous federated learning. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=EXnIyMVTL8s.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, 2025. doi: 10.1038/s41586-025-09442-9.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024. doi: 10.18653/v1/2024.findings-acl.33.

Soroosh Tayebi Arasteh, Alexander Ziller, Christiane Kuhl, Marcus Makowski, Sven Nebelung, Rickmer Braren, Daniel Rueckert, Daniel Truhn, and Georgios Kaissis. Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging. *Communications Medicine*, 4:46, 2024. doi: 10.1038/s43856-024-00462-6. URL https://www.nature.com/articles/s43856-024-00462-6.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8): 1930–1940, 2023. doi: 10.1038/s41591-023-02448-8.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025. doi: 10.1038/s41586-025-08866-7.

Ozlem Uzuner, Yuan Luo, and Peter Szolovits. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563, 2007. doi: 10.1197/jamia.M2444.

Guancheng Wan, Wenke Huang, and Mang Ye. Federated graph learning under domain shift with generalizable prototypes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15429–15437, 2024. doi: 10.1609/aaai.v38i14.29468.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024. doi: 10.1007/s11704-024-40231-1.

Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N. Ahmad Aziz, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862):265–270, 2021. doi: 10.1038/s41586-021-03583-3.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony Q. S. Quek, and H. Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020. doi: 10.1109/TIFS.2020.2988575.

Yipan Wei, Yuchen Zou, Hang Su, Yapeng Li, and Bo Du. FedProCL: Prototype-guided contrastive federated learning for medical image classification under concept shift. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4236–4239, 2025. doi: 10.1109/BIBM66473.2025.11356774.

Yipan Wei, Qikui Zhu, Zhendong Xia, Leyan Wang, Yuchen Zou, Yapeng Li, Yong Luo, and Bo Du. MedMerge: A training-free model merging framework for medical knowledge transfer into vision–language models. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2026. doi: 10.1109/TRPMS.2026.3662401.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2), 2025. doi: 10.1007/s11432-024-4222-0.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Y Zou, and Huaxiu Yao. MMed-RAG: Versatile multimodal RAG system for medical vision language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=s5epFPdIW6.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024. doi: 10.18653/v1/2024.findings-acl.372.

Xiyuan Yang, Wenke Huang, and Mang Ye. Dynamic personalized federated learning with adaptive differential privacy. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72181–72192, 2023. doi: 10.52202/075280-3160.

Xiyuan Yang, Wenke Huang, and Mang Ye. FedAS: Bridging inconsistency in personalized federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11986–11995, 2024. doi: 10.1109/CVPR52733.2024.01139.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=Q42f0dfjECO.

Jingyuan Zhang, Yiyang Duan, Shuaicheng Niu, Yang Cao, and Wei Yang Bryan Lim. Enhancing federated domain adaptation with multi-domain prototype-based federated fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=3wEGdrV5Cb.

A NOTATION

Table 3 collects the symbols used in the main text and in the appendices. The same symbol carries the same meaning throughout the paper.

Table 3: **The table defines notation for cases, hospitals, CD, EA, and evaluation metrics.**

Symbol	Meaning
Case, hospitals, and opinions	
x	raw originating case (the record that must stay local)
h_o	originating hospital; owns the case and the final decision
$j, \mathcal{H}$	index of a remote hospital; set of collaborating hospitals
R_j	private retrieval collection of hospital j
o_j	structured opinion returned by hospital j
y_j, D_j, r_j, u_j	primary diagnosis, differential diagnoses, rationale, and confidence inside o_j
y_{in}	incoming diagnosis at the originator before the decision guard
f_j	answering function of hospital j, so $o_j = f_j(P(x), R_j)$
$\hat{y}, \hat{y}_i^M$	emitted final primary diagnosis; the same for method M on case i
Controlled disclosure (CD)	
$P(x)$	reviewed packet released by CD
$\mathcal{R}$	raw-case field set (narrative, direct identifiers, record numbers, undeclared fields)
$\mathcal{K}, \Pi_{\mathcal{K}}(x)$	allowed field set of the packet schema; projection of x onto it
$\mathsf{Safe}(p, x)$	conjunction of schema, no-raw-text, and minimum-clinical-signal review predicates
$\Phi_{\text{disc}}, \Phi_{\text{collab}}, \Phi_{\text{adopt}}$	disclosure, collaboration, and adoption interfaces
Φ_{local}	local-only fallback diagnosis used when the disclosure gate rejects a packet
$\text{Fields}(\cdot)$, $\text{In}(v)$	fields of a message or message collection; fields observed by node v in one execution
F_{EA}	target system mapping case, packet, and opinions to the final diagnosis
Collaborative evidence processing (EA)	
$\mathcal{E}, \mathcal{C}_e, \mathcal{C}$	evidence channels; candidates proposed by channel e; candidate ledger
$q_e(c)$, $s_\theta(c)$	score of candidate c in channel e; frozen-ranker score of c with parameters θ
$n(c)$, $a(c)$	channels proposing c; channels in which c is the top-scoring candidate
c^*, $\Delta(c)$	top-ranked candidate; margin of a candidate over the runner-up
m, τ	minimum number of agreeing channels; promotion margin of the guard
$G(c)$	guard predicate: $c \neq y_{\text{in}}$, $n(c) \geq 1$, $a(c) \geq m$, $\Delta(c) \geq \tau$
Evaluation	
N, i	number of matched cases; case index
Y_i, ζ_i^M	accepted reference-label set of case i; correctness indicator of method M on case i
$A_b(M)$, A_{LR}	dataset-equal accuracy of method M on backbone b; accuracy of the LR reference
Acc, ECR	accuracy and error-correction rate (Appendix B.3); the regression rate is derived from them and is not reported separately
PER, CSR, N_{retained}	privacy exposure rate, clinical signal retention, and retained declared signals (Appendix D.6)

B THEORETICAL ANALYSIS

B.1 RAW-CASE NON-ACCESS UNDER CD

Let the raw case carry field set $\mathcal{R}$ (free-form narrative, direct identifiers, record numbers, undeclared fields) and let the packet schema use the disjoint allowed set $\mathcal{K}$, so $\mathcal{R} \cap \mathcal{K} = \emptyset$. With $p = \Pi_{\mathcal{K}}(x)$ and the gate of Equation (4),

$$P(x) = \begin{cases} p, & \mathsf{Safe}(p, x), \\ \perp, & \text{otherwise}, \end{cases} \tag{9}$$

and let $\text{In}(v)$ be the set of fields that node v observes over one execution.

Definition 1 (typed execution). An execution is *typed* if every message at every node equals a declared output of $\{\mathrm{Build}, \mathrm{Review}, \mathrm{Consult}, \mathrm{Coordinate}\}$ evaluated on declared inputs of that node, and no out-of-band channel exists.

Proposition 1 (raw-case non-access). In a typed execution, for every remote hospital j and for the coordinator c,

$$\mathrm{In}(j) \subseteq \mathcal{K} \cup \mathrm{Fields}(R_j), \qquad \mathrm{In}(c) \subseteq \mathcal{K} \cup \mathrm{Fields}\big(\{o_j\}\big), \tag{10}$$

hence $\mathcal{R} \cap \mathrm{In}(v) = \emptyset$ for every $v \neq h_o$.

Proof. By induction over the message order. (i) The only producer of a cross-boundary payload is $\mathrm{Review}\big(\mathrm{Build}(x)\big)$, whose declared output is $\Pi_{\mathcal{K}}(x)$ or $\perp$ by Equation (9), and $\mathrm{Fields}(\Pi_{\mathcal{K}}(x)) \subseteq \mathcal{K}$ by schema construction. (ii) The declared inputs of $\mathrm{Consult}$ are $\{P(x)\} \cup \{R_j\}$ and its declared output is the opinion $o_j = (y_j, D_j, r_j, u_j)$, a function of $P(x)$ and R_j alone. (iii) The declared inputs of $\mathrm{Coordinate}$ are $P(x)$ and $\{o_j\}$, and its declared output is a delivery record. Every message observed by $v \neq h_o$ is therefore a function of $\big(P(x), R_j\big)$, and since $\mathcal{R} \cap \mathcal{K} = \emptyset$ no raw field occurs in $\mathrm{In}(v)$. □

Corollary 1 (fail-closed). If $\neg\mathsf{Safe}(p, x)$ then $P(x) = \perp$, every remote input satisfies $\mathrm{In}(j) \subseteq \mathrm{Fields}(R_j)$, and the emitted diagnosis is

$$\hat{y} = \Phi_{\mathrm{local}}(x), \tag{11}$$

that is, no remote contribution enters the answer.

Proof. Immediate from Equation (9): with $P(x) = \perp$ the remote interface has no case payload, and the originator falls back to local diagnosis. □

Corollary 2 (identifier exposure). Let $\mathcal{D} \subseteq \mathcal{R}$ be the direct-identifier fields. Then $\mathcal{D} \cap \mathrm{In}(v) = \emptyset$ for all $v \neq h_o$, so the released representation carries no direct identifier and the exposure rate defined in Appendix D.6 is zero by construction:

$$\mathrm{PER}\big(P(x)\big) = 0. \tag{12}$$

B.2 Selection soundness under EA

With the ledger $\mathcal{C}$, the provenance counts $n(c)$ and $a(c)$ of Equation (5), and the frozen ranker s_θ, define

$$\Delta(c) = s_\theta(c) - \max_{c' \in \mathcal{C} \setminus \{c\}} s_\theta(c'), \qquad \max \emptyset := -\infty, \tag{13}$$

and the guard predicate

$$G(c) := \big[c \neq y_{\mathrm{in}}\big] \wedge \big[n(c) \geq 1\big] \wedge \big[a(c) \geq m\big] \wedge \big[\Delta(c) \geq \tau\big]. \tag{14}$$

Selection follows Equation (6): with $c^* = \arg\max_{c \in \mathcal{C}} s_\theta(c)$,

$$\hat{y} = \begin{cases} c^*, & G(c^*) = 1, \\ y_{\mathrm{in}}, & G(c^*) = 0. \end{cases} \tag{15}$$

Lemma 1 (output range). $\hat{y} \in \mathcal{C} \cup \{y_{\mathrm{in}}\}$.

Proof. By cases on $G(c^*)$ in Equation (15): the first branch emits $c^* \in \mathcal{C}$, the second emits y_{in}. □

Proposition 2 (promotion soundness). If $\hat{y} \neq y_{\mathrm{in}}$ then $G(\hat{y}) = 1$; equivalently, every emitted override satisfies

$$n(\hat{y}) \geq 1, \qquad a(\hat{y}) \geq m, \qquad \Delta(\hat{y}) \geq \tau. \tag{16}$$

Proof. $\hat{y} \neq y_{\mathrm{in}}$ holds only in the first branch of Equation (15), which requires $G(c^*) = 1$ with $\hat{y} = c^*$; expanding G by Equation (14) gives the three inequalities. □

Corollary 3 (no uncorroborated promotion). For any $c \in \mathcal{C}$ with $a(c) < m$, $\Pr[\hat{y} = c] = 0$ unless $c = y_{\text{in}}$.

Proof. By Proposition 2, $\hat{y} = c \neq y_{\text{in}}$ implies $a(c) \geq m$. □

Proposition 3 (monotone restriction). Let E_m be the promotion event $\{\hat{y} \neq y_{\text{in}}\}$ under threshold m. For $m' \geq m$,

$$E_{m'} \subseteq E_m, \tag{17}$$

and the accuracy effect of raising the threshold is confined to $E_m \setminus E_{m'}$.

Proof. The predicate G is monotone decreasing in m: if $a(c) \geq m'$ then $a(c) \geq m$, hence $G_{m'}(c) = 1 \Rightarrow G_m(c) = 1$ for every c, so $E_{m'} = \{G_{m'}(c^*) = 1\} \subseteq \{G_m(c^*) = 1\} = E_m$. Cases in $E_m \setminus E_{m'}$ keep $\hat{y} = y_{\text{in}}$ under m', so only they can change accuracy. □

Proposition 4 (retention and conditional correctness). Let Y be the accepted diagnosis set.

$$G(c^*) = 0 \implies \hat{y} = y_{\text{in}}, \tag{18}$$

$$G(c^*) = 1 \ \wedge \ c^* \in Y \implies \hat{y} \in Y, \tag{19}$$

$$G(c^*) = 0 \ \wedge \ y_{\text{in}} \in Y \implies \hat{y} \in Y. \tag{20}$$

Proof. The first line restates the second branch of Equation (15). For the second, $G(c^*) = 1$ gives $\hat{y} = c^* \in Y$. For the third, $G(c^*) = 0$ gives $\hat{y} = y_{\text{in}} \in Y$. □

Corollary 4 (error containment). A correct incoming diagnosis can be lost only by a promotion of a candidate outside Y, which by Proposition 2 requires $a(c^*) \geq m$ and $\Delta(c^*) \geq \tau$; consequently

$$\Pr[\hat{y} \notin Y, \ y_{\text{in}} \in Y] \ \leq \ \Pr\big[G(c^*) = 1, \ c^* \notin Y\big]. \tag{21}$$

B.3 Co-primary metrics: Acc and ECR

For a matched cohort of N cases, let $\zeta_i^M = \mathbb{1}[\hat{y}_i^M \in Y_i]$ and $A_{\text{LR}} = \text{Acc(LR)}$, and define

$$\text{ECR}(M) = \Pr(\zeta^M = 1 \mid \zeta^{\text{LR}} = 0), \qquad \text{RGR}(M) = \Pr(\zeta^M = 0 \mid \zeta^{\text{LR}} = 1). \tag{22}$$

ECR is defined only when LR makes at least one error, and is then zero for LR by construction; RGR is N/A when LR has no correct case.

Only Acc and ECR are reported as metrics. Once A_{LR} and ECR are fixed, Equation (23) determines RGR as well, so RGR is a derived quantity used to interpret the two reported metrics rather than a third number tabulated separately.

Proposition 5 (exact decomposition). For any method evaluated on the same cases as its LR reference with $0 < A_{\text{LR}} < 1$,

$$\text{Acc}(M) = A_{\text{LR}} + (1 - A_{\text{LR}})\,\text{ECR}(M) - A_{\text{LR}}\,\text{RGR}(M). \tag{23}$$

Proof. Partition the cohort by LR correctness. Among the $A_{\text{LR}}N$ locally correct cases, M retains $(1 - \text{RGR}(M))A_{\text{LR}}N$; among the $(1 - A_{\text{LR}})N$ local errors it recovers $\text{ECR}(M)(1 - A_{\text{LR}})N$. The two correct sets are disjoint, so summing and dividing by N gives Equation (23). □

Corollary 5 (balance condition).

$$\text{Acc}(M) = A_{\text{LR}} \iff (1 - A_{\text{LR}})\,\text{ECR}(M) = A_{\text{LR}}\,\text{RGR}(M). \tag{24}$$

A method that recovers local errors therefore only breaks even on accuracy once it overwrites an $A_{\text{LR}}/(1 - A_{\text{LR}})$ fraction of them in locally correct cases.

Proof. Set the right-hand side of Equation (23) equal to A_{LR} and rearrange. □

Corollary 6 (differential form and weights). For variations $\Delta\,\mathrm{ECR}$ and $\Delta\,\mathrm{RGR}$,

$$\Delta\,\mathrm{Acc} = (1 - A_{\mathrm{LR}})\,\Delta\,\mathrm{ECR} - A_{\mathrm{LR}}\,\Delta\,\mathrm{RGR}, \tag{25}$$

$$\frac{\partial\,\mathrm{Acc}}{\partial\,\mathrm{ECR}} = 1 - A_{\mathrm{LR}}, \qquad \frac{\partial\,\mathrm{Acc}}{\partial\,\mathrm{RGR}} = -A_{\mathrm{LR}}. \tag{26}$$

The weights $1 - A_{\mathrm{LR}}$ and A_{LR} are complementary, so a single accuracy value cannot separate recovery from regression unless both metrics are reported.

Proof. Differentiate Equation (23) with respect to ECR and RGR; the coefficients are $1 - A_{\mathrm{LR}}$ and $-A_{\mathrm{LR}}$. □

C Implementation Details

This section groups the setup of the study: datasets (Appendix C.1), compared methods (Appendix C.2), and our own implementation (Appendix C.3).

C.1 Dataset Details

Table 4 records the six datasets and their scales. The two large Difficult datasets are first reduced by deterministic label-stratified 1% sampling. A paper-level disjoint-cohort filter then removes every test case seen during policy development, so the retained test cohort differs from the source cohort; for seeds 42/48/51 it contains 63/63/63, 71/71/71, 292/294/293, 115/115/115, 148/148/148, and 16/17/16 cases of S2D, K-S2D, KDSD, K-Sym, Ext., and MAC-302, that is, 705/708/706 cases and 2,119 cases per backbone over the three seeds. Table 5 summarizes the input representation and sampling protocol without treating the resulting hospital nodes as real institutions.

Table 4: **The table reports each dataset's group, source size, and label count.**

Dataset	Group	Source cases	Labels
Symptom-to-Diagnosis (S2D)	Simple	1,065	22
Kaggle Symptom2Disease (K-S2D)	Simple	1,200	24
Kaggle Disease-Symptom Description (KDSD)	Simple	4,920	41
Kaggle Symptoms-to-Diseases (K-Sym)	Difficult	192,715	254
Disease-Symptom Extensive (Ext.)	Difficult	246,945	773
MAC-302	Rare	298	297

Each panel below is one verbatim case per dataset (truncated where the record is long), that is, what a case looks like to the models.

S2D: free-form patient symptom narrative

"I've been having a lot of pain in my neck and back. I've also been having trouble with my balance and coordination. I've been coughing a lot and my limbs feel weak."

Diagnosis: cervical spondylosis

K-S2D: free-form symptom narrative

"I have been experiencing a skin rash on my arms, legs, and torso for the past few weeks. It is red, itchy, and covered in dry, scaly patches."

Diagnosis: Psoriasis

KDSD: compact symptom-description record

"Symptoms: itching, skin rash, nodal skin eruptions, dischromic patches"

Diagnosis: Fungal infection

K-Sym: compact symptom list

"Symptoms: anxiety and nervousness, shortness of breath, depressive or psychotic symptoms, chest tightness, palpitations, irregular heartbeat, breathing fast"

Diagnosis: panic disorder

Ext.: compact symptom list

"Symptoms: difficulty speaking, throat swelling, throat feels tight"

Diagnosis: vocal cord polyp

MAC-302: long-form clinical case narrative

"Follow-up Presentation: Basic Information: Female proband, born full term at 37th week, with a birth weight of 2 Kg and a head circumference of 32 cm. Parents are healthy and had a consanguineous marriage. Clinical Presentation: The affected child has features typical of Wiedemann–Rautenstrauch syndrome . . . "

Diagnosis: Trisomy 18

Table 5: **The table specifies model-facing inputs and sampling for each dataset.** Source-cohort selection precedes the common partitioning protocol.

	Data protocol	
Dataset	**Input form**	**Source sampling**
S2D	Free-form patient symptom narrative	Complete declared source cohort
K-S2D	Free-form symptom narrative	Complete declared source cohort
KDSD	Compact symptom-description record	Complete declared source cohort
K-Sym	Compact symptom list	Deterministic label-stratified 1%
Ext.	Compact symptom list	Deterministic label-stratified 1%
MAC-302	Long-form clinical case narrative	Complete declared source cohort

Common partitioning protocol. For every retained dataset cohort, 60% of cases form the RAG pool and 40% form the experiment pool. Each pool is allocated across three simulated hospital nodes according to a 50/30/20 capability gradient. Within every node, the experiment allocation is split directly into 70% train, 15% validation, and 15% test. The canonical evaluation repeats this deterministic procedure with seeds 42, 48, and 51. All aggregate results average datasets equally rather than weighting them by test size.

C.2 Compared Methods

Table 6 distinguishes what each evaluated method can access, how collaboration proceeds, and where the final decision is made. Every method predicts before gold labels are opened.

C.3 Our Implementation

Shared settings. We evaluate DeepSeek V4 Flash (DeepSeek, 2026), Qwen 3.7 Flash (Qwen, 2026), and GPT-5.6-Luna (OpenAI, 2026). Every hospital within a run uses the same base LLM and temperature 0. Node-local retrieval uses precomputed Qwen3-Embedding-4B representations (Qwen Team, 2025; Karpukhin et al., 2020; Khattab & Zaharia, 2020) and returns the top five documents. Model-facing stages use schema-validated outputs. In particular, a `DiagnosticOpinion` contains a primary diagnosis, differential, rationale, and confidence. Private-RAG evidence is represented separately by bounded node identifier, retrieval rank, candidate label, and score records.

Support RAG follows medical retrieval benchmarks (Xiong et al., 2024) and can be federated across sites (Joy & Su, 2026); medical knowledge can also be supplied as graphs or ontologies (Hogan et al., 2021; Köhler et al., 2021) or injected by weight-space merging (Wei et al., 2026), and multi-view evidence can be fused with uncertainty-aware attention (Li et al., 2024a) or aligned to external semantic names (Li et al., 2024b). Because retrieved text can carry identifiers, the packet review checks plaintext directly (Kim et al., 2023).

CD packet and local review contract. CD transforms the raw originator case into a closed-schema packet before any cross-hospital call. Table 7 groups the packet fields by their operational role. Free-form raw narrative, direct identifiers, record numbers, and undeclared fields are not packet fields.

The local reviewer revalidates the closed schema, filters raw-case fields and matching plaintext from the source record, and requires at least one controlled clinical concept or abnormal finding. A failed check blocks release and invokes the local-only fallback. Every decision records the passed checks,

Table 6: **The table compares evidence access, collaboration protocol, and final decision authority across methods.**

	Information exchange		Decision authority
Method	**Case/evidence access**	**Collaboration protocol**	**Aggregation or decision**
LR	Originator raw case and local RAG	No remote collaboration	Originator emits one structured diagnosis (Lewis et al., 2020)
RV	Every hospital receives raw case context and its local RAG	Independent structured opinions; inspired by multi-agent debate but uses no debate rounds (Du et al., 2024)	Majority vote; confidence and lexical order break ties
PV	Every hospital receives the reviewed packet and its local RAG	Same opinion-and-vote pattern as RV, but with packet-mediated context	Majority vote under the RV rule
MAC	Every hospital receives raw case context and its local RAG	Initial opinions followed by up to two synchronous discussion rounds	Originator model performs supervisor synthesis (Chen et al., 2025b)
MedChain	Referred hospital receives raw case context and its local RAG	Referral, history, examination, diagnosis, and treatment stages; at most two referrals	Final stage returns the structured diagnosis (Liu et al., 2025a)
MedECHO (Ours)	Originator retains raw case; remote hospitals receive reviewed packet and private RAG only	All schema-valid returned opinions are relayed with source provenance	Originator fuses evidence, manages candidates, applies decision guard, and emits final diagnosis

Table 7: **The table lists the representation and release constraint for each CD packet field group.**

	Validated release contract	
Field group	**Representation**	**Release constraint**
Case reference	SHA-256 case identifier and coarse origin tag	No original case or patient identifier
Demographics	Ten-year age band, sex category, and coarse region code	Controlled values only
Clinical signal	Up to six chief-complaint concepts, bounded concept/finding lists, and discretized vital signs	Allowlisted concepts or validated medical terminology codes
Context	Coarse imaging, disease-course, risk-factor, and suspected-specialty codes	Enumerated values; no free-form narrative

fallback reason, and hashed case identifier. For MedECHO, each remote hospital queries its private R_j using only the reviewed packet. The coordinator receives the reviewed packet and schema-valid structured opinions for stable-order delivery and audit. Raw x remains at the originator, where evidence fusion and diagnosis take place.

EA execution contract. For each released packet, EA executes the following auditable sequence:

1. eligible remote hospitals retrieve from their own R_j using packet text and return one schema-valid `DiagnosticOpinion`; failed calls are logged and omitted;
2. the coordinator validates delivery and relays all returned structured opinions in stable hospital order;
3. the structured opinions and retrieval evidence return to the originator, where an initial collaborative diagnosis is formed;
4. the originator builds a label-free candidate ledger from the local opinion, sealed remote opinions, bounded textual mentions, HPO evidence (Köhler et al., 2021), and the read-only Support corpus; and
5. the frozen decision stack applies retrieval-aware scoring and the source-agreement guard. A candidate failing the margin or source checks leaves the diagnosis unchanged.

Gold labels are not inputs to these five prediction steps, and the coordinator never emits the final diagnosis.

Decision stack. The stack has two stages. *Retrieval-aware selection* produces the incoming diagnosis y_{in} from the ledger and the retrieval evidence, without consulting the ranker below. *The ranker and guard* then test whether some other ledger candidate should replace y_{in}. The ranker scores every candidate with a frozen centroid model over a 60-dimensional label-free feature vector: 28 features carried over from the v3 candidate schema (source membership and rank reciprocals, retrieval and confidence maxima, source-type, node, and consensus counts, with the raw local confidence replaced by interacted terms) and 32 features added by the ranker (per-source maximum scores and top-1 margins for the existing Top-20, Support RAG, HPO, and bounded-mention channels, textual overlap and duplicate-span statistics, and interactions between source agreement and the local reading). Vectors are standardized by a stored mean and scale, and the score is the distance to a negative centroid minus the distance to a positive centroid, both fitted on 60 development cases; candidate identity, case identity, and labels are excluded. Let c^* be the top-scoring candidate and $\Delta(c^*)$ its margin over the runner-up. The guard promotes c^* only if it is not already the pre-selection reading, the source agreement satisfies $a(c^*) \geq m$, the margin satisfies $\Delta(c^*) \geq \tau$, and retrieval support is available; otherwise y_{in} stands. In the deployed policy $m = 2$ and $\tau = 0.15$, so the guard replaces the pre-selection reading only on a corroborated, clearly separated candidate, and the guarded decision is the final diagnosis.

Table 8: **The table maps available information, explicit exclusions, and outputs at each system stage.**

	Information-flow contract		
Stage	**Available information**	**Explicitly excluded**	**Output**
Originating hospital	Raw case, local RAG, local opinion	None within local boundary	Reviewed collaboration packet
Remote hospital	Approved packet, remote private RAG	Raw case and originator-local records	Structured diagnostic opinion
Coordinator	Reviewed packet and structured opinions	Raw case, raw retrieval documents, clinical scoring, diagnosis authority	Valid opinion delivery records
Originator decision	Local evidence and returned structured opinions	Remote raw records and retrieval collections	Final primary diagnosis and differential

D Supplementary Experiments

D.1 Dataset-Level Main Results

Tables 9 and 10 provide the exact values visualized by Figure 4. They use the same matched three-seed cohorts, Acc/ECR definitions, and dataset-equal aggregation as the main results.

D.2 Candidate Adoption Rate

The adoption rate asks how often a method that has the accepted diagnosis in its candidate set also emits it as the primary diagnosis; we compute it as the dataset-equal strict Top-1 rate divided by the dataset-equal disease-found rate, over the same six datasets and three seeds. Figure 10 reports it for every method and LLM. MedECHO converts 90.11% of candidates it finds into the final diagnosis on DeepSeek and 88.26% on Qwen and GPT, while no baseline exceeds 69.35%, 65.05%, or 69.67%.

D.3 Sequential System Comparison

RV, PV, and MedECHO can also be read as three complete communication-and-decision contracts rather than as modules (their access, collaboration, and decision fields are in Table 6): RV votes over opinions formed from the raw case, PV applies the same vote to opinions formed from the reviewed packet, and MedECHO keeps the packet but hands the decision to the frozen originator

Table 9: **MEDECHO leads both Acc and ECR on every Simple dataset.** Acc and ECR (%) as mean±sample standard deviation over seeds 42/48/51; Avg weights datasets equally and bold marks the best value within each base LLM.

Methods	S2D		K-S2D		KDSD		Avg	
	Acc	ECR	Acc	ECR	Acc	ECR	Acc	ECR
DeepSeek V4 Flash								
LR [NeurIPS 2020]	74.07±6.42	0.00±0.00	67.61±1.41	0.00±0.00	72.36±1.99	0.00±0.00	71.35±3.21	0.00±0.00
RV [ICML 2024]	70.90±8.74	5.56±9.62	67.61±2.82	8.64±4.26	73.04±1.96	3.67±1.98	70.52±3.73	5.96±1.76
PV	57.67±15.00	5.56±9.62	56.81±5.33	7.25±5.02	68.03±0.97	20.13±3.25	60.84±6.04	10.98±2.13
MAC [npj Digital Medicine 2025]	70.37±9.57	10.85±5.44	70.42±1.41	18.68±5.79	77.94±3.70	27.93±7.74	72.91±3.74	19.15±2.95
MedChain [NeurIPS 2025]	64.02±9.57	17.52±20.92	60.09±1.63	15.90±1.99	61.77±0.42	7.84±3.01	61.96±2.57	13.75±6.44
MedECHO (Ours)	**80.95±5.72**	**51.80±5.85**	**83.57±2.15**	**59.43±1.30**	**99.89±0.20**	**99.61±0.68**	**88.14±1.44**	**70.28±1.92**
Qwen 3.7 Flash								
LR [NeurIPS 2020]	65.08±9.91	0.00±0.00	63.38±1.41	0.00±0.00	70.65±4.39	0.00±0.00	66.37±5.04	0.00±0.00
RV [ICML 2024]	65.08±8.84	7.47±0.84	63.85±0.81	3.85±0.15	70.65±4.23	2.24±0.83	66.53±4.60	4.52±0.44
PV	50.26±12.93	8.06±10.55	53.52±0.00	10.27±2.27	64.16±2.17	22.49±4.08	55.98±3.65	13.61±2.83
MAC [npj Digital Medicine 2025]	65.61±4.85	17.16±2.63	64.79±1.41	7.65±3.64	70.19±2.87	11.99±3.72	66.86±2.96	12.27±2.56
MedChain [NeurIPS 2025]	61.90±5.50	15.93±4.49	60.09±2.15	9.03±2.57	64.73±2.72	4.72±0.72	62.24±3.41	9.89±2.45
MedECHO (Ours)	**78.84±5.10**	**53.89±12.51**	**80.28±3.73**	**53.65±8.12**	**99.89±0.20**	**99.61±0.67**	**86.33±0.73**	**69.05±2.77**
GPT-5.6-Luna								
LR [NeurIPS 2020]	71.43±7.27	0.00±0.00	65.26±3.54	0.00±0.00	81.23±1.75	0.00±0.00	72.64±2.63	0.00±0.00
RV [ICML 2024]	73.02±6.35	11.49±10.80	63.85±8.25	4.55±7.87	80.78±1.00	3.54±1.52	72.55±3.32	6.53±1.18
PV	59.26±7.16	6.21±7.32	53.52±3.73	9.30±5.41	72.81±1.70	29.71±6.63	61.86±2.61	15.08±4.14
MAC [npj Digital Medicine 2025]	70.37±8.74	12.93±13.56	62.44±3.54	11.20±6.06	82.14±1.63	19.22±9.19	71.65±3.45	14.45±2.13
MedChain [NeurIPS 2025]	66.67±6.92	18.21±15.57	63.38±2.82	13.20±4.83	68.83±1.11	5.39±1.49	66.29±2.72	12.27±6.52
MedECHO (Ours)	**81.48±3.30**	**51.19±4.84**	**80.75±2.15**	**51.47±3.29**	**99.89±0.20**	**99.44±0.98**	**87.37±0.68**	**67.36±0.77**

Table 10: **MEDECHO leads both Acc and ECR on the Difficult and Rare datasets.** Acc and ECR (%) as mean±sample standard deviation over seeds 42/48/51; Avg weights datasets equally and bold marks the best value within each base LLM.

Methods	K-Sym		Ext.		MAC-302		Avg	
	Acc	ECR	Acc	ECR	Acc	ECR	Acc	ECR
DeepSeek V4 Flash								
LR [NeurIPS 2020]	28.99±2.66	0.00±0.00	18.92±2.70	0.00±0.00	18.63±27.33	0.00±0.00	22.18±10.03	0.00±0.00
RV [ICML 2024]	**35.07±7.29**	16.84±6.31	19.82±0.39	8.86±1.44	28.80±23.91	12.50±0.00	27.90±10.51	12.73±2.07
PV	26.96±5.22	2.08±2.65	16.67±0.39	0.82±0.83	16.54±23.74	0.00±0.00	20.06±9.63	0.97±1.15
MAC [npj Digital Medicine 2025]	27.25±1.33	4.88±1.10	18.69±2.73	4.75±2.15	24.51±18.77	8.33±9.55	23.48±6.52	5.99±2.83
MedChain [NeurIPS 2025]	22.32±7.03	11.51±3.49	16.44±1.70	10.00±0.72	42.65±10.89	27.08±25.26	27.14±5.66	16.20±7.86
MedECHO (Ours)	**35.07±3.05**	**31.50±3.02**	**40.32±1.41**	**41.13±0.99**	**51.10±10.01**	**35.42±32.07**	**42.16±2.37**	**36.02±9.78**
Qwen 3.7 Flash								
LR [NeurIPS 2020]	26.38±1.33	0.00±0.00	17.79±2.81	0.00±0.00	36.76±6.38	0.00±0.00	26.98±1.92	0.00±0.00
RV [ICML 2024]	28.41±7.24	13.32±6.68	18.24±3.10	8.69±3.85	38.73±6.60	6.73±5.92	28.46±4.59	9.58±4.32
PV	23.19±4.11	3.50±5.06	15.77±1.41	0.54±0.94	34.80±7.80	0.00±0.00	24.59±4.07	1.35±2.00
MAC [npj Digital Medicine 2025]	27.25±0.50	9.03±2.60	17.79±0.78	4.07±2.09	36.64±10.16	10.44±11.17	27.23±3.72	7.84±4.63
MedChain [NeurIPS 2025]	23.48±3.48	12.59±2.38	16.89±2.03	8.47±1.00	42.77±4.85	16.50±6.73	27.71±2.43	12.52±2.32
MedECHO (Ours)	**35.94±3.51**	**31.04±6.06**	**36.26±2.37**	**36.42±1.78**	**49.02±12.61**	**23.23±8.75**	**40.41±5.11**	**30.23±3.58**
GPT-5.6-Luna								
LR [NeurIPS 2020]	31.59±2.66	0.00±0.00	19.59±3.58	0.00±0.00	32.72±9.80	0.00±0.00	27.97±2.51	0.00±0.00
RV [ICML 2024]	33.33±8.26	13.33±7.64	22.52±4.13	10.60±4.42	36.76±1.27	8.33±8.33	30.87±1.60	10.76±4.00
PV	29.28±3.92	3.84±2.65	18.69±3.19	0.84±0.81	24.63±11.16	0.00±0.00	24.20±2.23	1.56±1.05
MAC [npj Digital Medicine 2025]	32.75±2.80	6.74±1.66	18.47±2.81	3.04±2.30	30.51±5.19	5.56±4.81	27.25±2.88	5.11±1.55
MedChain [NeurIPS 2025]	27.83±7.97	14.98±5.79	18.69±2.17	11.51±1.52	**38.85±4.39**	**12.04±4.24**	28.46±0.93	12.84±2.38
MedECHO (Ours)	**34.49±2.66**	**30.59±3.29**	**38.29±1.95**	**38.13±1.73**	30.76±7.00	6.48±5.78	**34.51±2.61**	**25.07±1.29**

stack. Over the six-dataset cohort and three seeds, that contract change alone accounts for the gap: dataset-equal Acc/ECR is 49.21/9.34 (RV), 40.45/5.97 (PV), and 65.15/53.15 (MedECHO) on DeepSeek, 47.49/7.05, 40.28/7.48, and 63.37/49.64 on Qwen, and 51.71/8.64, 43.03/8.32, and 60.94/46.22 on GPT. Tables 9 and 10 give the per-dataset values behind these averages, and the staged replays below decompose the MedECHO column.

D.4 AGENT CAPABILITY, MODEL HETEROGENEITY, AND WORKLOAD

Table 1 in Section 4.3 reports all seed-42 dataset-level composition results. The workload comparison additionally uses cache-only replay. Cross-family MedECHO uses 502,674 normalized Tokens and 538 logical calls against 1.55M Tokens and 1,526 calls for MAC and 1.57M Tokens and 1,615 calls for MedChain. Calls and Tokens are reconstructed by cache-only replay with one fixed tokenizer, so these are workload counts rather than cost or latency measurements.

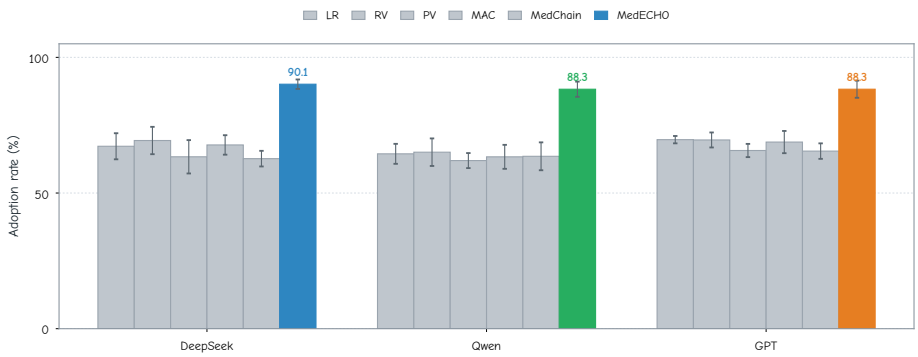

Figure 10: **The plot compares candidate adoption rates across methods and base LLMs.** Adoption is dataset-equal strict Top-1 rate divided by dataset-equal candidate-found rate; bars show three-seed means with sample SD and color marks MEDECHO.

D.5 EA ABLATION IN DETAIL

Table 2 of the main text builds the originator decision stack stage by stage, so this subsection instead removes the individual evidence sources from the full system. Only the auxiliary retrieval corpus carries a large load: removing Support RAG drops dataset-equal Acc to 54.59/55.22% on DeepSeek/Qwen against 66.73/66.90% for the full system, whereas removing the bounded-mention or the HPO channel leaves accuracy within noise (67.10/68.01% and 66.73/66.90%). Support RAG, bounded mentions, and HPO similarity are system settings rather than EA steps: they change which evidence enters the candidate set, not how the originator compares and adopts it.

D.6 CD DISCLOSURE AUDIT DEFINITIONS

Privacy Exposure Rate is

$$\mathrm{PER} = \frac{\mathrm{IdentifierLeakRate} + \mathrm{SensitiveKeyExposureRate}}{2}, \tag{27}$$

where lower is better, and Clinical Signal Retention is

$$\mathrm{CSR} = \frac{N_{\mathrm{retained}}}{13}, \tag{28}$$

where higher is better. The deterministic canaries measure the implemented schema, plaintext, and signal-retention contract; Figure 5 in the main text reports the per-strategy comparison.

D.7 WORKFLOW DIAGNOSTICS AND ORIGINATOR CORRECTION

The staged view of candidate discovery and final decision is in Figure 8 of the main text. Candidate found measures whether the accepted diagnosis appears in the available candidate set, whereas Acc measures the final primary diagnosis; under packet voting the candidate-found rates are 63.46/64.92/65.54% against final Top-1 Acc values of 40.45/40.28/43.03%, and originator re-evaluation raises MedECHO Acc to 65.15/63.37/60.94%.

Across the sealed replay the same stage decides the outcome: the accepted diagnosis is in most cases already present in the candidate set, and the originator-side promotion step is what turns it into the final answer. This matches the candidate-to-decision analysis of Section 3.2: MedECHO constructs a usable candidate set, and EA's selection rule is the step that commits to it.

D.8 ILLUSTRATIVE CASE COMPARISON

Figure 9 in the main text shows the node opinions and originator ledger for this case. LR, PV, and MAC answer sickle cell crisis; MedChain answers spondylosis. These additional outputs belong to the same illustrative case and do not estimate a population-level effect.