About Me



NAME

Pretom Roy Ovi

TITLE

Assistant Professor, Dept of Data Science

YEARS AT UNT

2 years

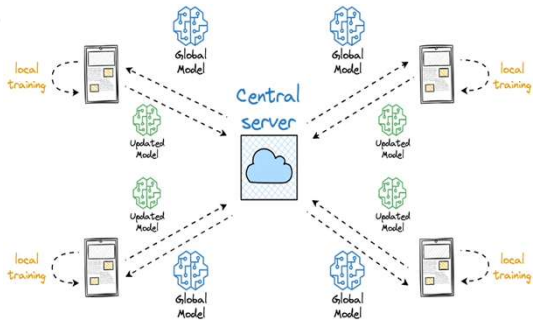
RESEARCH FOCUS AREAS

- Privacy Enhanced Machine Learning
- Federated Learning
- Large Language Modeling
- Data Modeling

An Unlearning Framework for Targeted Knowledge Removal.

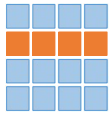
What is it FL?

Decentralized training a shared model without sharing raw data.



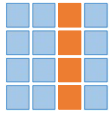
What can be Unlearn?

Grid rows = clients, columns = classes



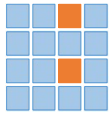
Client Wise

Removes the entire contribution of a specific client from the trained global model, ensuring that its previously learned information no longer influences the model.



Class Wise

Removes the learned knowledge of a specific class across all participating clients while preserving the model's ability to recognize other classes.



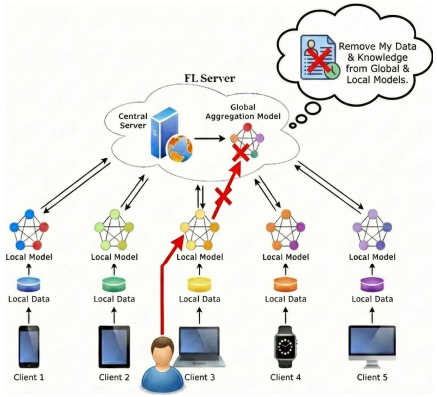
Client–Class Wise

Selectively removes the influence of a particular class contributed by specific clients, while preserving knowledge of the same class learned from other clients.

Why do we need it?

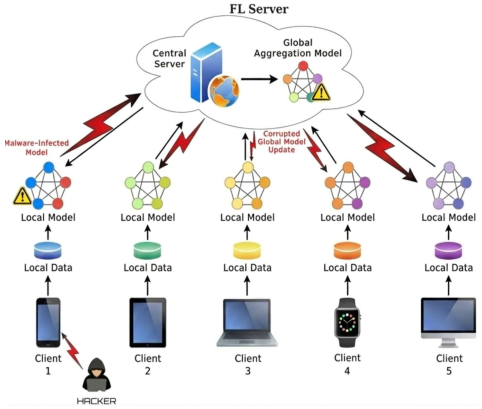
Privacy & deletion requests

Even after a client deletes its local data or leaves the federation, its previously learned information remains embedded in the global model, creating privacy risks and challenges for the right to be forgotten.



Corrupted / poisoned data

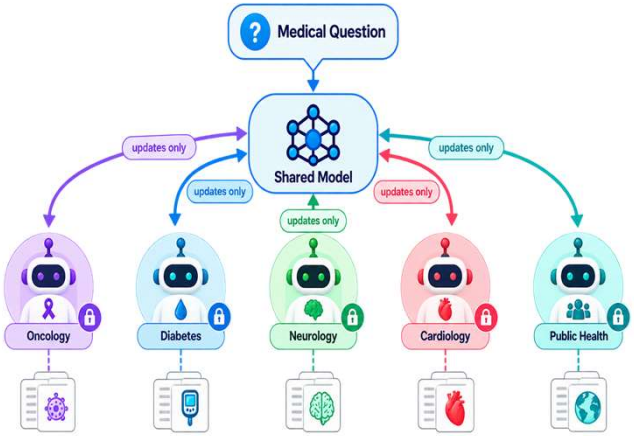
Malicious or faulty clients can introduce poisoned, mislabeled, or corrupted data during training. Their harmful influence remains in the global model, potentially compromising its accuracy, reliability, and security.



A Multi-Agent AI Framework for Biomedical Q/A in Healthcare

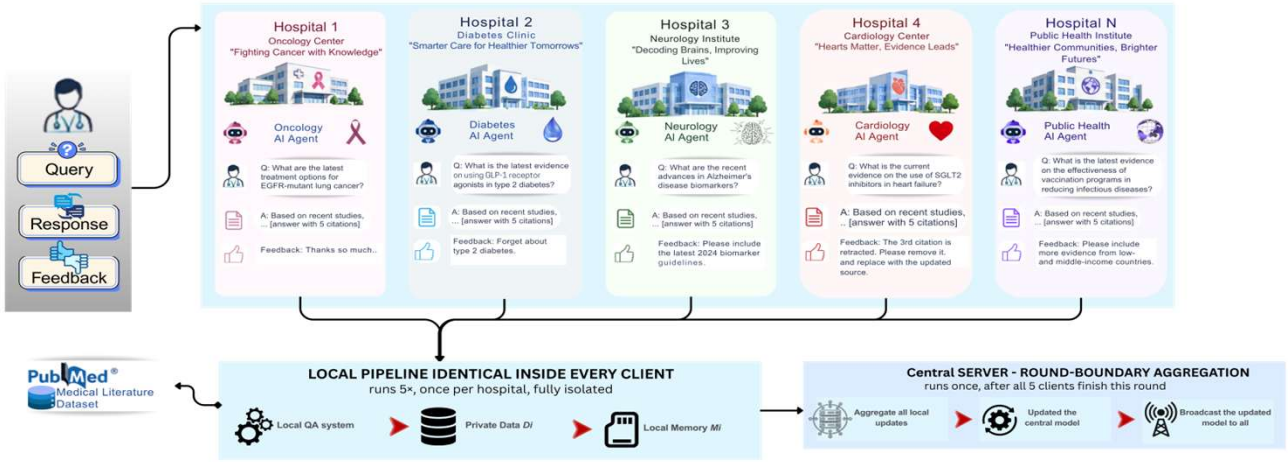
What is Multi-Agent Biomedical QA?

Multiple specialized AI agents answer medical questions using relevant scientific evidence while keeping their local information private.



What did we do?

We proposed a Privacy-Preserving Multi-Agent Biomedical Question-Answering Framework that improves citation quality through expert feedback and synchronizes on-demand topic suppression across agents



Examples of Feedback Signals

- Remove outdated or retracted citations
- Prefer newer or higher-quality evidence
- Add missing subtopics or perspectives
- Improve ranking for specific clinical guidelines
- Include diverse populations (e.g., LMIC)
- General relevance and correctness feedback

Result at Next Round

- ✓ Improved and more relevant answers
- ✓ Updated and higher-quality citations
- ✓ Changes apply to all clients
- ✓ Continuous learning without sharing data
- ✓ Better clinical decision support globally

EthosAI: Ethical and Culturally Aware AI Tutor

THE DESIGN CHALLENGE

A response can sound convincing and still be poorly grounded, too advanced, unfairly framed, or unsuitable for the learner.

Retrieval can help with evidence, but it does not by itself govern every educational risk.

Grounding

Can claims be traced to course evidence?

Learner fit

Does the explanation suit this learner?

Safeguards

Should the system answer, warn or block?

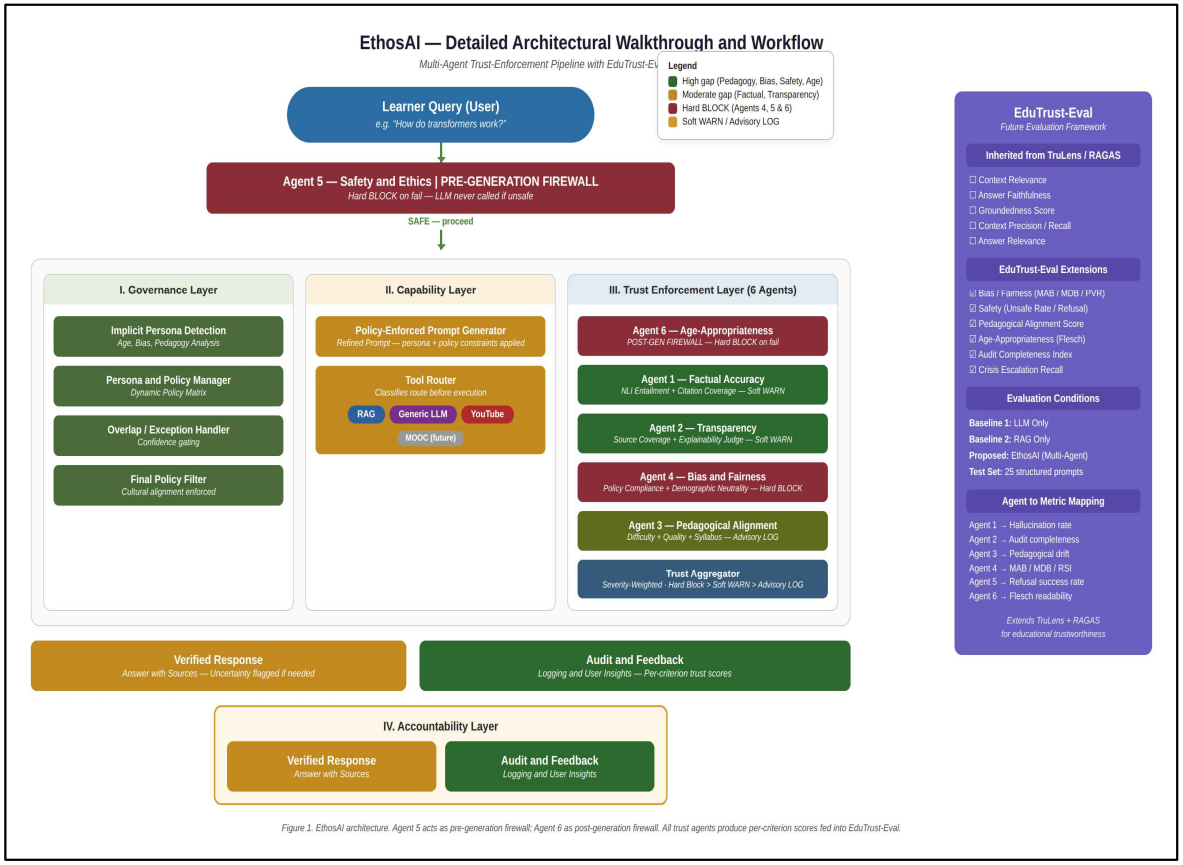


Figure 1. EthosAI architecture. Agent 5 acts as pre-generation firewall; Agent 6 as post-generation firewall. All trust agents produce per-criterion scores fed into EduTrust-Eval.

Trustworthiness at the core: Six agent actions

Each agent checks a different risk in the response and records what it finds.

QUERY + COURSE EVIDENCE → DRAFT ANSWER → SIX AGENT ACTIONS → AGGREGATED VERDICT → DELIVERY / BLOCK

AGENT 1 • FACTUAL ACCURACY

CHECKS
- Claims supported by course sources
- Citations and contradictions

FINDING • PASS

Claims were supported by retrieved course context and linked to specific pages.

AGENT 2 • TRANSPARENCY

CHECKS
- Source traceability
- Explanation of evidence
- Audit completeness

FINDING • PASS

The answer was traceable to course pages, with a clear source trail.

AGENT 3 • PEDAGOGICAL ALIGNMENT

CHECKS
- Learner-level fit
- Teaching clarity and quality
- Syllabus alignment

FINDING • WARN

Formal probability notation was introduced without explaining it clearly enough for this learner.

AGENT 4 • BIAS & FAIRNESS

CHECKS
- Policy compliance
- Neutrality and potential bias
- Appropriate difficulty

FINDING • PASS

No policy-compliance or demographic-neutrality issue was identified.

AGENT 5 • SAFETY & ETHICS

CHECKS
- Harmful or unsafe content
- Crisis signals
- Dishonesty and jailbreak patterns

FINDING • PASS

This definition request showed no unsafe or academic-dishonesty pattern. Other cases can be blocked before generation.

AGENT 6 • AGE-APPROPRIATENESS

CHECKS
- Readability and language complexity
- Cognitive load
- Developmental suitability

FINDING • PASS

The answer was age-suitable overall, though the notation was a mild point of difficulty.

OVERALL OUTCOME • PASS - response delivered; a pedagogical advisory was logged for educator review.

Safety contrast: an academic-dishonesty pattern can trigger a hard block before the LLM call; other agents are then not invoked.