M.Sc. Thesis Defense

Erfan Nemati

A non-academic tech guy in software industry

Technical Product Manager, Software Engineer

Startups & Enterprise

Bachelor's degree, Computer Engineering, IAUEC

Master's degree, Artificial Intelligence, IAUEC

Data

GDPR, Regulatory, Dark web, Privacy, Data breach, Trust, Data sharing

M.Sc. Thesis Defense
Islamic Azad University, E-Campus
Faculty of Engineering

A Blockchain-Based Recommender System for Enhanced Security & Privacy in Data-Oriented Organizations

Author: Erfan Nemati

Thesis Supervisor: Dr. Parisa Rahmani

Consulting Advisor: Dr. Parvaneh Asghari

Thesis Examiner: Dr. Maryam Taajobian

Summer 2025

Problem Statement

Data Sharing

C

Centralized

A single point where all data is stored, making it vulnerable to data breaches and misuse.

P

Privacy & Trust

A core conflict where users are concerned about how their personal data is collected, stored, and used

T

Transparency

The absence of a clear process for how organizations use user data, which erodes user confidence

M

Monetization

Organizations face significant challenges in securely monetizing their valuable data

Research Significance

Industry Growth

AI Disruption

Economic Incentive

Digital Transformation

Valuation
$5.3 billion by 2030

Regulatory and GDPR Pressure

Research Gaps

Production

Operational

H — Holistic

Focuses on a single component, like a specific algorithm or an incentive model

D — Efficient Design

Clear design patterns to effectively manage the complex trade-offs between privacy, accuracy, and performance.

E — Economic

Reliable, stable models for data sharing and monetization in a B2B marketplace is limited.

I — Integration

fail to fully and synergistically integrate all the vital components—privacy, security, traceability, incentives, and governance—into a single, practical system.

Research Questions

Role Abilities

A — Architecture

How combine BC and FL to enhance security and privacy while managing scalability and costs?

R — Role of Blockchain

What specific functions can be delegated to the Blockchain and smart contracts to improve security, transparency, trust, and traceability?

P — Privacy

Which privacy-preserving techniques can be combined with the proposed architecture to provide a measurable level of privacy against common threats?

P — Performance

What is the performance overhead caused by the Blockchain and encryption components, and is it acceptable for practical use by data-driven organizations?

$ — Monetization

How can a secure model for inter-organizational data sharing be designed using Blockchain, and what kind of revenue model can incentivize participation?

T — Trade-offs

What is the nature of the trade-off between privacy, recommendation accuracy, system performance, and cost in the proposed architecture?

Research Hypotheses

H1 — Auditability

The blockchain layer will significantly increase auditability and transparency.

H2 — Privacy-Accuracy

There will be a measurable privacy-accuracy trade-off.

H3 — Performance

Blockchain integration will introduce a manageable performance overhead.

H4 — Secure

The proposed architecture and its economic model will enable secure data sharing

H5 — Incentivize

The proposed architecture and its economic model will enable secure incentivize participation.

Improvement

Literature Review

Recommender Systems

RS

What is RS

A recommender system is a software subclass of information filtering systems that predicts a user's preferences for an item. These systems analyze large amounts of data to provide personalized suggestions in various fields like e-commerce, entertainment, and social media. Their success depends on access to rich, high-quality user data.

Collaborative Filtering (CF)

01

User-Item interaction
Similar Past to Similar Future

Content-Based Filtering (CB)

02

Item-based
User’s profile of interests

Neural Matrix Factorization (NeuMF)

03

Find simple and hidden pattern in combination of GMF and Multi-layer perceptrons

Blockchain

What is BC

Blockchain is a distributed, shared, and immutable ledger that securely and transparently records transactions without the need for a central authority. Its key features include decentralization, immutability, and controlled transparency.

Public Blockchains

01

Decentralized

Permissioned Blockchains

02

Enterprise & B2B

Federated Learning

What is FL

Federated Learning is a distributed machine learning approach where multiple clients, such as mobile devices or organizations, collaboratively train a global model. This is done without sharing their raw, local data with a central server or with each other. Instead, only model updates, such as gradients or weights, are sent to a central server for aggregation.

Centralized FL

01

Federated Averaging (FedAvg)

03

Algorithm
Initialization, Local Training, Update
Communication, Global Aggregation

Decentralized FL

03

Structured Process

Design Science Research

Knowledge Generation

Artifact Creation

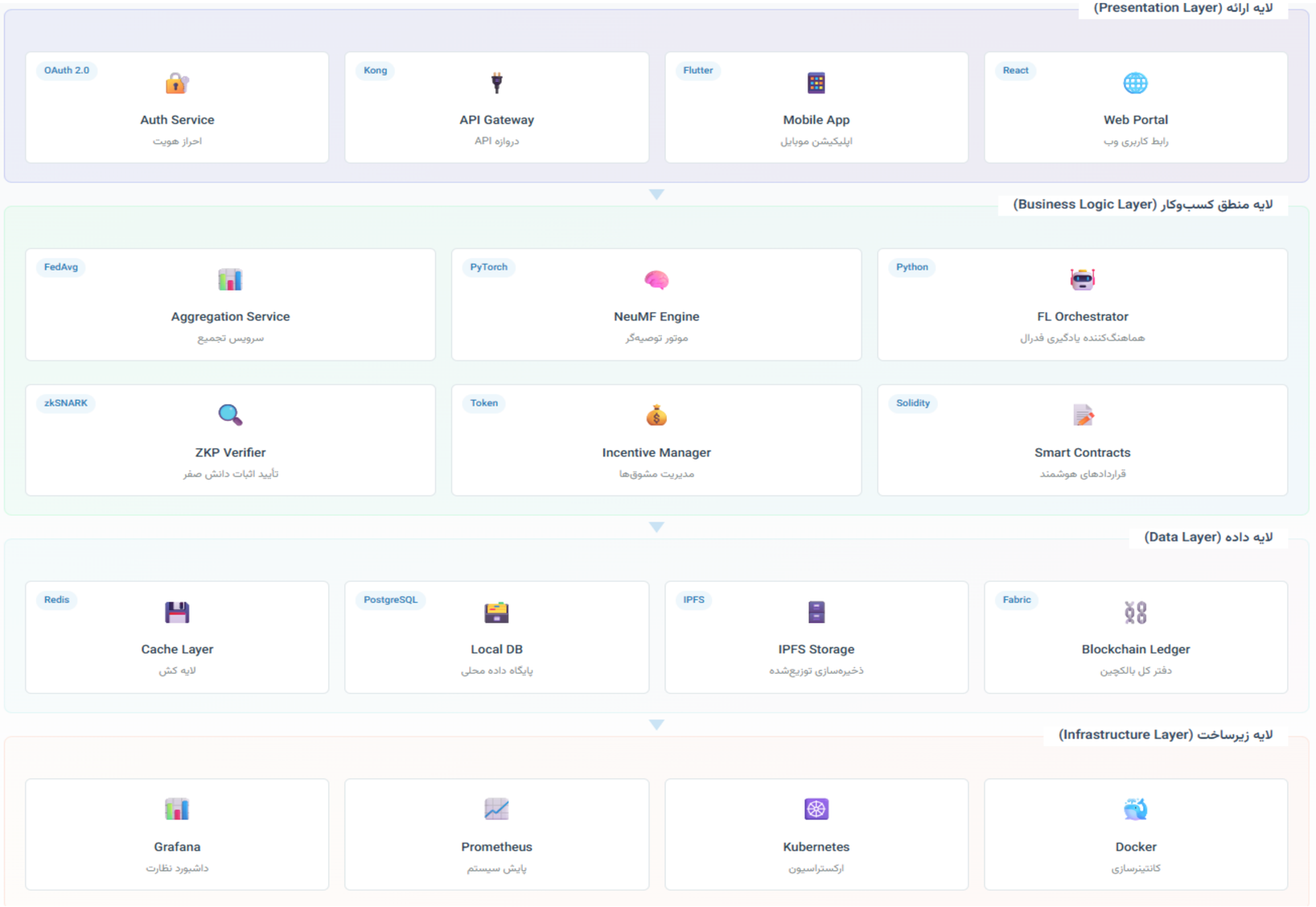
Problem Solving Focus

create a generalizable solution

Architecture Overview

A Multi-Layered Approach

The proposed architecture is a multi-layered, hybrid solution designed to create a secure and scalable data marketplace. It strategically divides tasks among three main layers to optimize performance and security.



How it works?
The Process.

01

Consortium Onboarding

A data-providing organization joins the network by registering its decentralized identity (DID) and verifiable credentials (VCs) on the blockchain. This establishes a trusted on-chain identity.

02

Asset Registration

The data-providing organization stores its datasets and models in a secure off-chain storage system (like IPFS) and registers their cryptographic hash (CID) on the blockchain. This guarantees data integrity without exposing the data itself.

03

Request for Training

A data-seeking organization (a "client") submits a request to the system to train a specific recommender model, including its budget and requirements.

How it works?
The Process.

04

Reverse Auction

A smart contract initiates a reverse auction to select the most suitable data providers for the task based on criteria like data quality and cost. This is an incentive mechanism to attract quality contributors.

05

Local Model Training

The selected organizations download the global model and train it locally on their private data. The raw data never leaves the organization's secure perimeter.

06

Proof Generation

After local training, each organization generates a Zero-Knowledge Proof (ZKP). This proof cryptographically verifies that the training was conducted correctly and on valid data, without revealing the underlying information.

How it works?
The Process.

07

On-chain Verification

Organizations send their model updates and the ZKP to the blockchain. A smart contract verifies the ZKP to ensure the updates are valid and not malicious. Only valid updates are accepted for aggregation.

08

Global Aggregation

An aggregation server, or a smart contract in a decentralized setup, combines the verified local model updates to create a new, improved global model.

09

Reward Distribution

Once the final model is ready, a smart contract automatically distributes the payment or token-based rewards to the organizations whose contributions were successfully verified on-chain. This provides a transparent and automated incentive system.

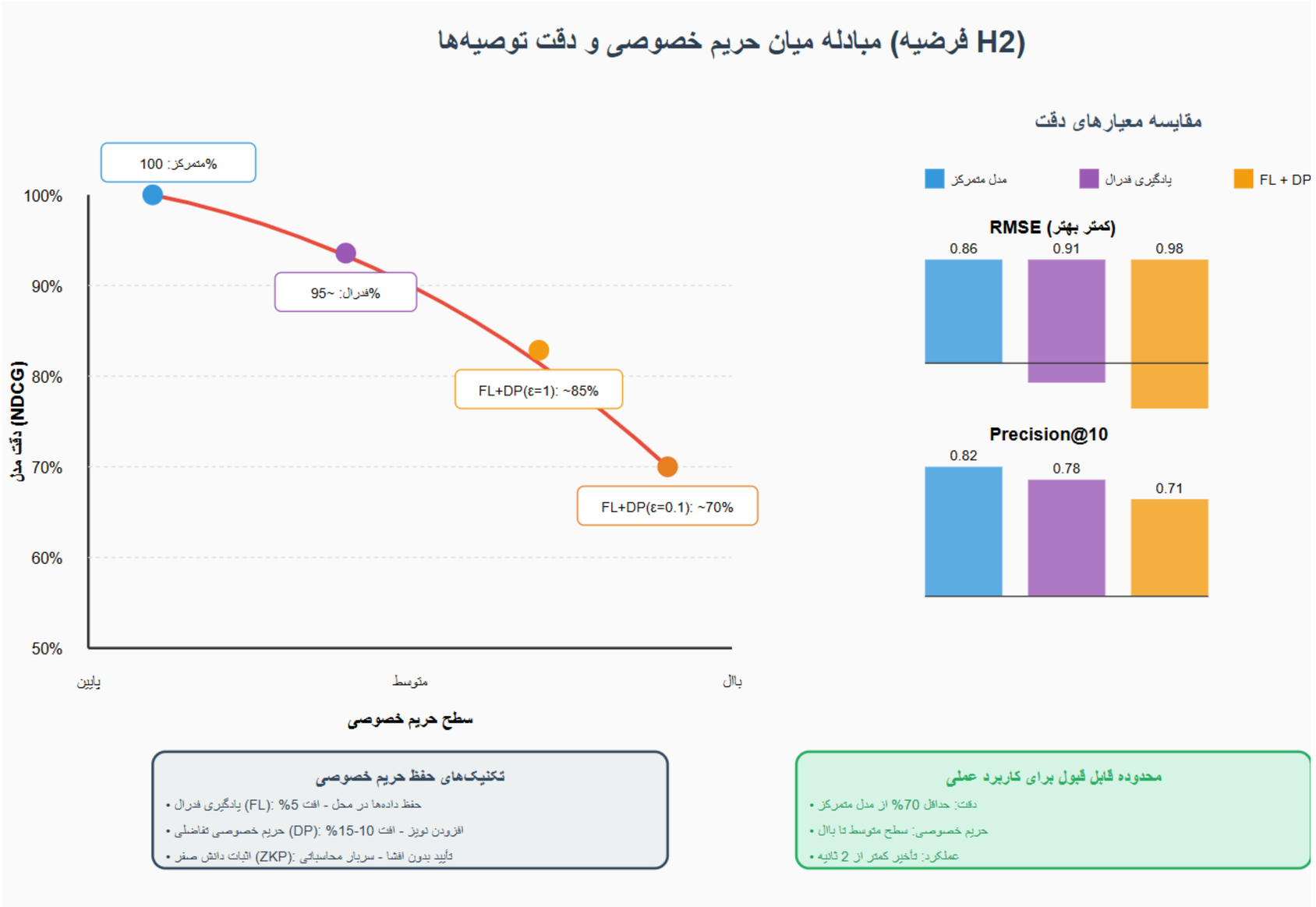
H1-Auditability

The H1 hypothesis is fully confirmed. Our architecture uses a permissioned blockchain to create a tamper-proof audit trail of all critical system events. This blockchain is not just a passive log; it's an active enforcement engine. Smart contracts automatically verify every transaction and model update. This means that the system's security and transparency are guaranteed by transparent, verifiable code, not by a single, trusted central authority. The architecture ensures that any malicious behavior is permanently recorded and

Results & Evaluation

H2- Privacy-Accuracy

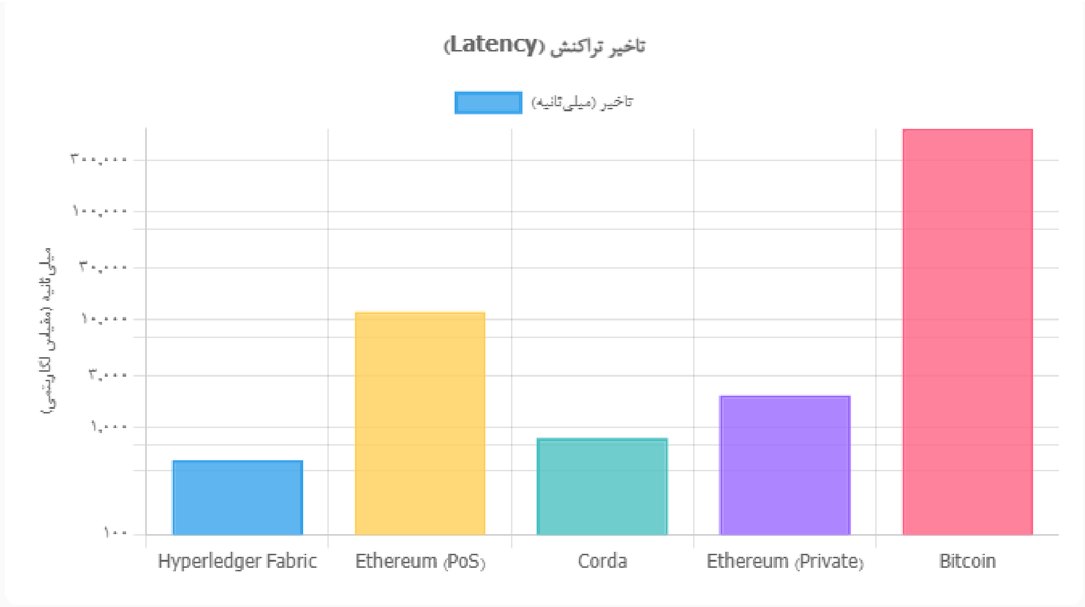
The H2 hypothesis is confirmed: there's an inherent trade-off between privacy and accuracy. Shifting from a centralized model to a federated one, which protects privacy by keeping raw data local, leads to a measurable drop in accuracy. For example, studies on the MovieLens dataset show a 4.8% reduction in the NDCG@10 metric for a federated model compared to a centralized one. This demonstrates that privacy comes at a cost, and adding further privacy measures like Differential Privacy would likely increase this accuracy loss.

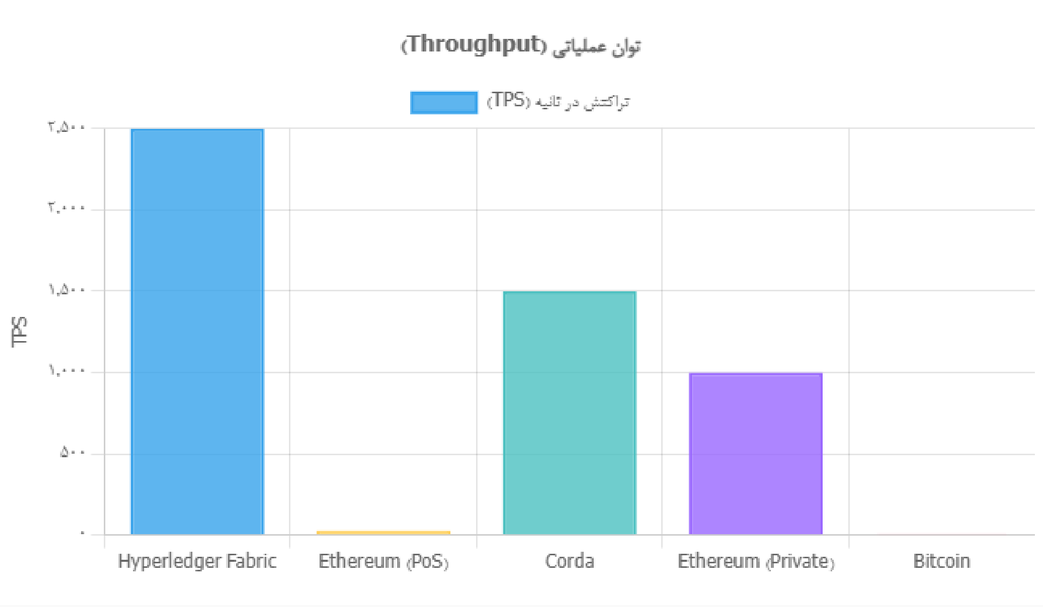


Results & Evaluation

H3-Performance

Confirmed. the performance overhead from blockchain operations is manageable and acceptable for this use case. This is achieved through a strategic on-chain/off-chain design. The architecture avoids storing large datasets and models directly on the blockchain, which would cause significant latency and cost. Instead, the blockchain is used for lightweight, high-value tasks like logging, identity management, and verifying proofs of computation. This design ensures the blockchain doesn't become a bottleneck. Your thesis cites benchmarks for Hyperledger Fabric, which show it can handle between 1,500 and 2,900 transactions per second (TPS) with low latency, far exceeding the needs of a collaborative model-training application.

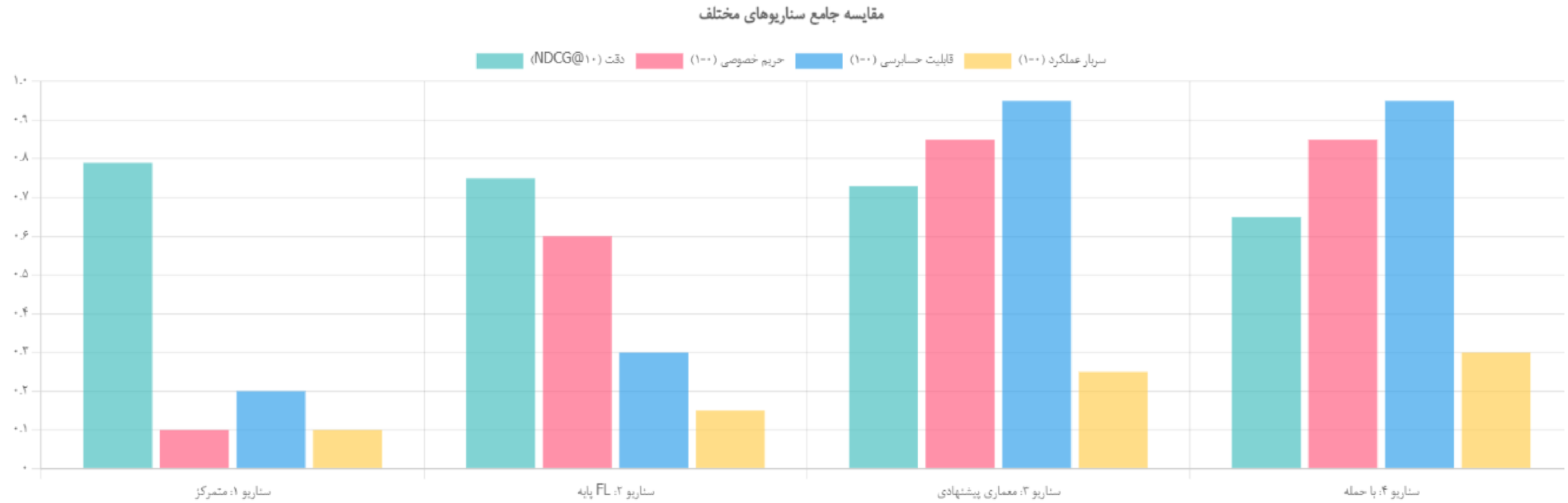




Results & Evaluation

H4 & H5 -Business Model Validation

H4 (Secure Data Sharing): Confirmed. The use of private channels in Hyperledger Fabric allows competing organizations to collaborate securely on a specific project, solving a key business problem.
H5 (Incentives): Confirmed. The smart contract-managed "verifiable contribution-to-automated reward" loop creates a transparent and self-enforcing economic incentive, encouraging honest participation.



Conclusion

Future Work

- Smart contract evaluation imrovement
- **Economic Model**
- **Advanced FL Algorithm**
- **Real world Evaluation**
- ZKP Improvement

Special Thanks.

Q&A.