REALM

www.realm-ai.eu

A Blockchain-based Approach for Model Card Accountability and Regulatory Compliance

International Workshop on Blockchain for Information Systems (BC4IS24) and Blockchain for Trusted Data Sharing (B4TDS)

(Co-located with the 36th International Conference on Advanced Information Systems Engineering (CAiSE) 2024)

Ankur Lohachab, Visara Urovi

IDS, DACS, Maastricht University

03 June 2024

Maastricht University

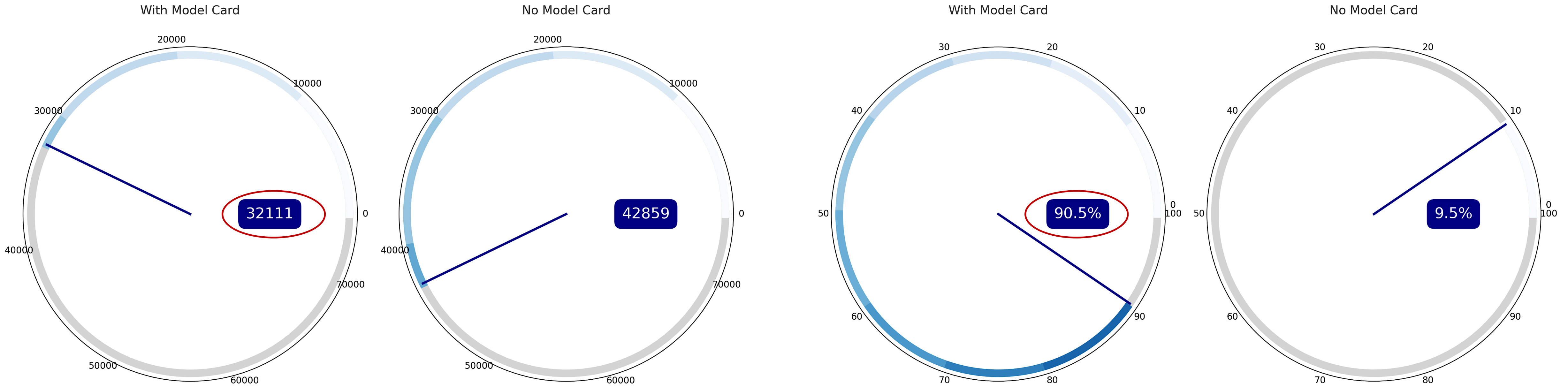
Funded by
the European Union

Outline

- □ Introduction
  - Context
  - Model Cards
  - Problematics
- □ Our Contribution
  - Proposed Approach
  - Roles and Interactions
  - Model Card Lifecycle
- □ Key Takeaways
  - Transparency
  - Traceability
  - Accountability
- □ Conclusion and Perspectives

Introduction (1/7)

Context



(a) Model Card Adoption

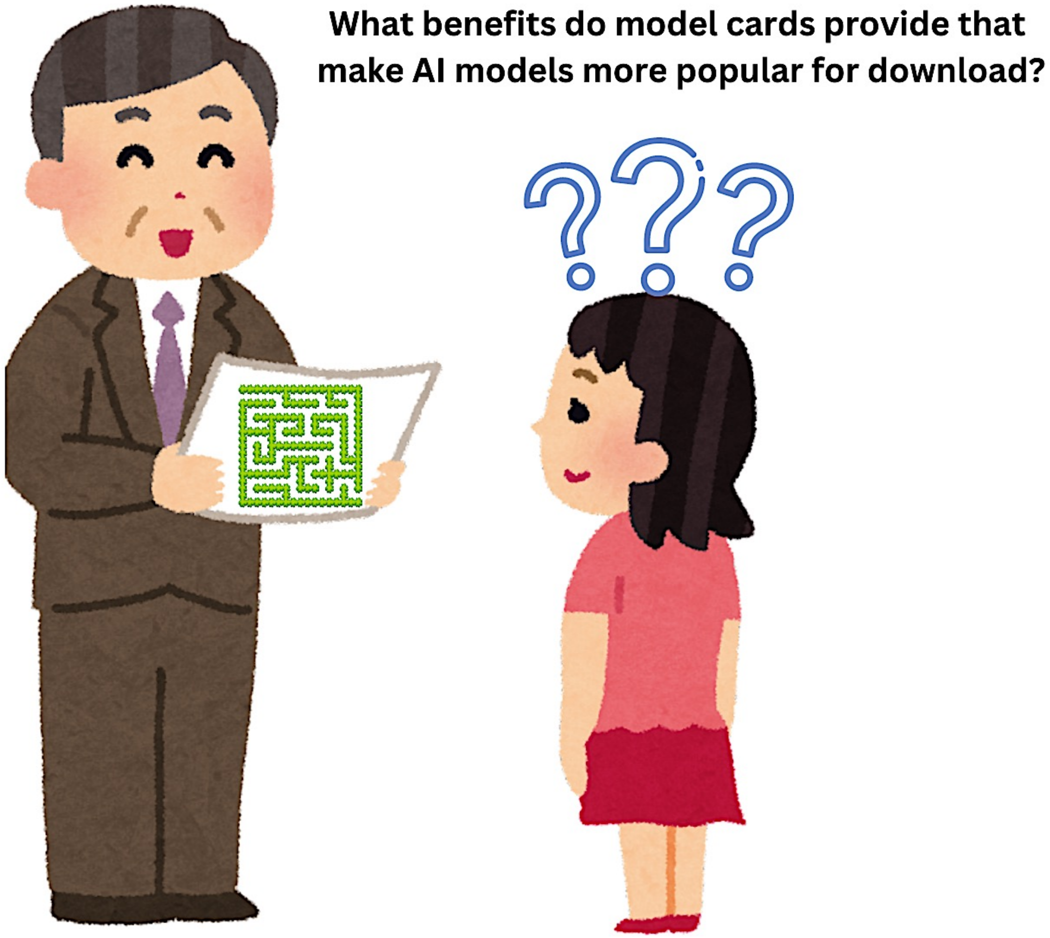
(b) Downloaded Traffic

Figure 1: 42.8% of repositories have model cards, yet they account for 90.5% of total downloads

(Data Source: Liang et al.[2]*)*

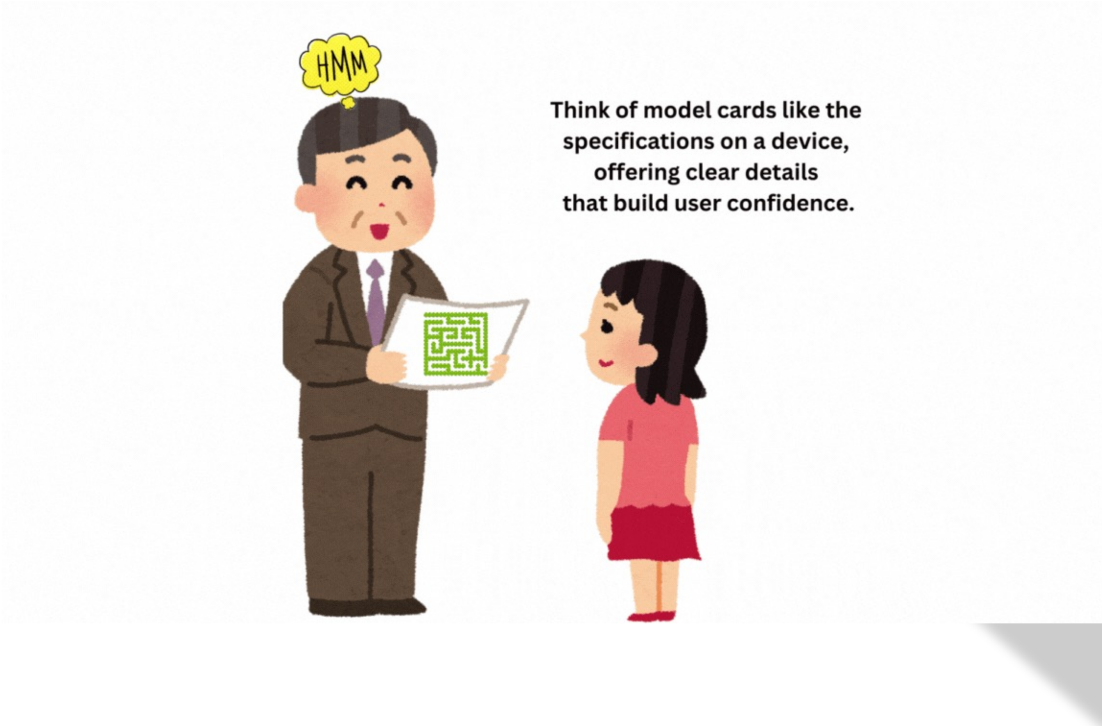
Introduction (2/7)

Context


What benefits do model cards provide that make AI models more popular for download?
???

Introduction (3/7)

Context



Think of model cards like the specifications on a device, offering clear details that build user confidence.

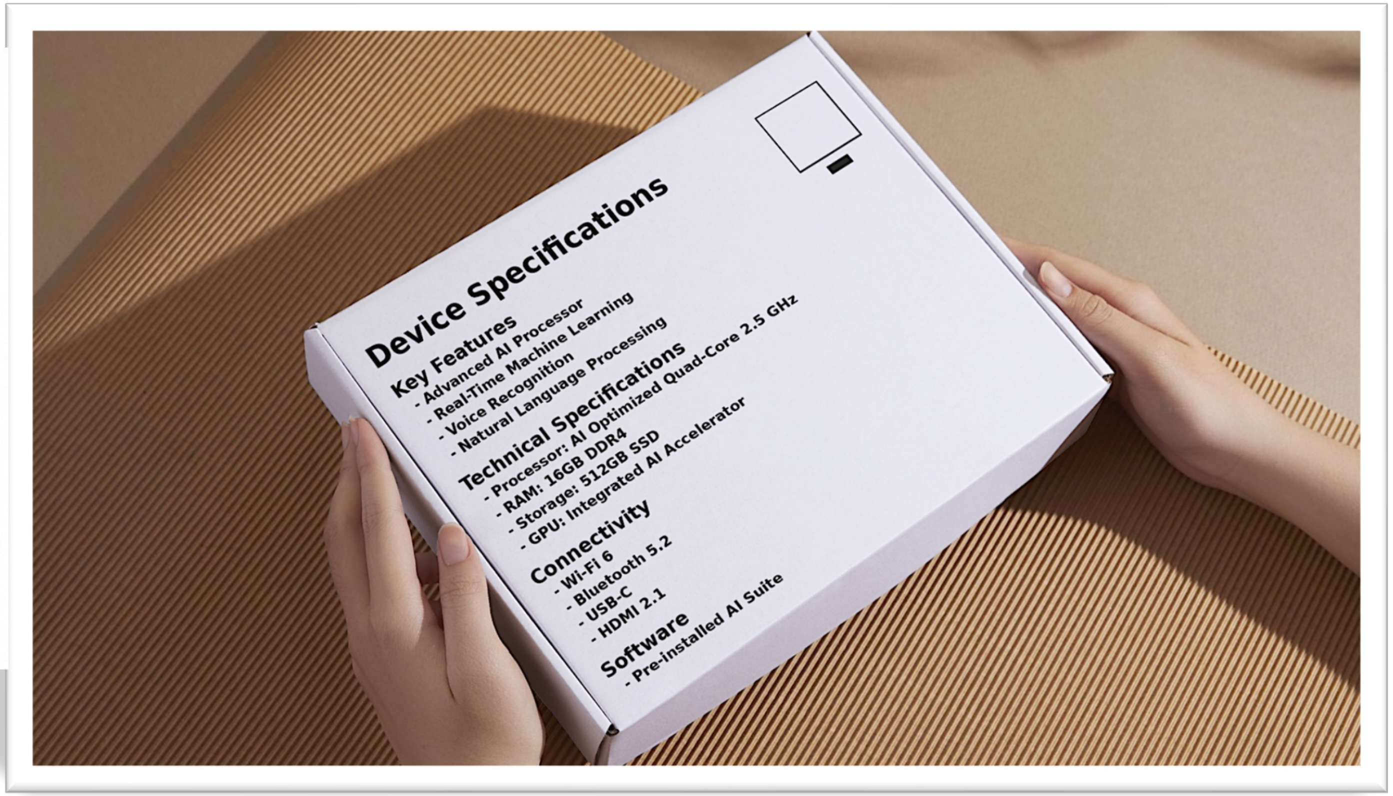


Figure 2: Model Card provides information at a glance, explaining your device

Introduction (4/7)

Motivating Scenario

1. **Clinical Transparency**

   □ Provides detailed information on how AI models make decisions, helping clinicians understand and trust the AI's recommendations.

2. **Patient Safety**

   □ Documents potential risks and failure modes of AI models, ensuring safe implementation in patient care.

3. **Regulatory Compliance**

   □ Helps meet healthcare regulatory standards by providing comprehensive documentation of model development, validation, and deployment processes.

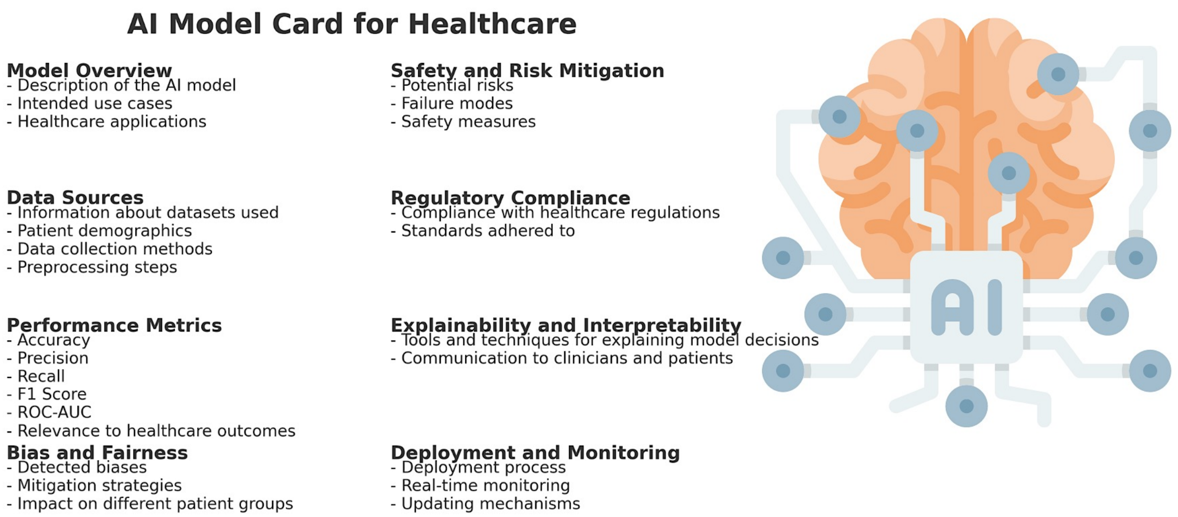
4. **Bias Identification**

   □ Highlights any biases in the training data or model behavior, promoting fair and equitable treatment across diverse patient populations.

5. **Performance Metrics**

   □ Clearly outlines the performance metrics of AI models, including accuracy, sensitivity, and specificity, tailored to specific healthcare applications.

How Can Model Cards for AI Models in Healthcare Provide Benefits and What Are Their Key Applications?



AI Model Card for Healthcare

Model Overview
- Description of the AI model
- Intended use cases
- Healthcare applications

Data Sources
- Information about datasets used
- Patient demographics
- Data collection methods
- Preprocessing steps

Performance Metrics
- Accuracy
- Precision
- Recall
- F1 Score
- ROC-AUC
- Relevance to healthcare outcomes

Bias and Fairness
- Detected biases
- Mitigation strategies
- Impact on different patient groups

Safety and Risk Mitigation
- Potential risks
- Failure modes
- Safety measures

Regulatory Compliance
- Compliance with healthcare regulations
- Standards adhered to

Explainability and Interpretability
- Tools and techniques for explaining model decisions
- Communication to clinicians and patients

Deployment and Monitoring
- Deployment process
- Real-time monitoring
- Updating mechanisms

Figure 3: A Snippet of AI Model Card Template for Healthcare

Introduction (5/7)

Model Cards

Table 1: Simplified Comparison of a Few Methods to Create Model Cards

Method	Standardization	Customization	Integration	Ethical Considerations
Model Card Toolkit (MCT)[3]	Provides a consistent template that includes sections for model details, performance, fairness, and limitations. Utilizes Jinja templates for customization and is integrated with TFX.	Moderate customization; allows adding extra sections but within a predefined structure. Supports JSON Schema for custom sections.	Seamless integration with TensorFlow Extended (TFX) and other ML frameworks. Exports model cards in various formats such as HTML, PDF, and Markdown.	High; includes specific sections for ethical considerations and bias analysis. Facilitates fairness evaluation with quantitative metrics and visualizations.
Hugging Face Model Cards[4]	High; provides a standard template widely recognized in the community. Structured format includes sections for intended uses, limitations, biases, and evaluation metrics.	Moderate; allows for some customization within the template. Supports additional fields and markdown for detailed descriptions and explanations.	High; tightly integrated with the Hugging Face model repository. Models and their cards are co-located, making it easy to find documentation alongside models.	High; encourages documentation of ethical considerations and fairness analysis. Provides tools for bias detection and mitigation. Model cards can include detailed ethical analyses and usage recommendations.

Introduction (6/7)

Problematics

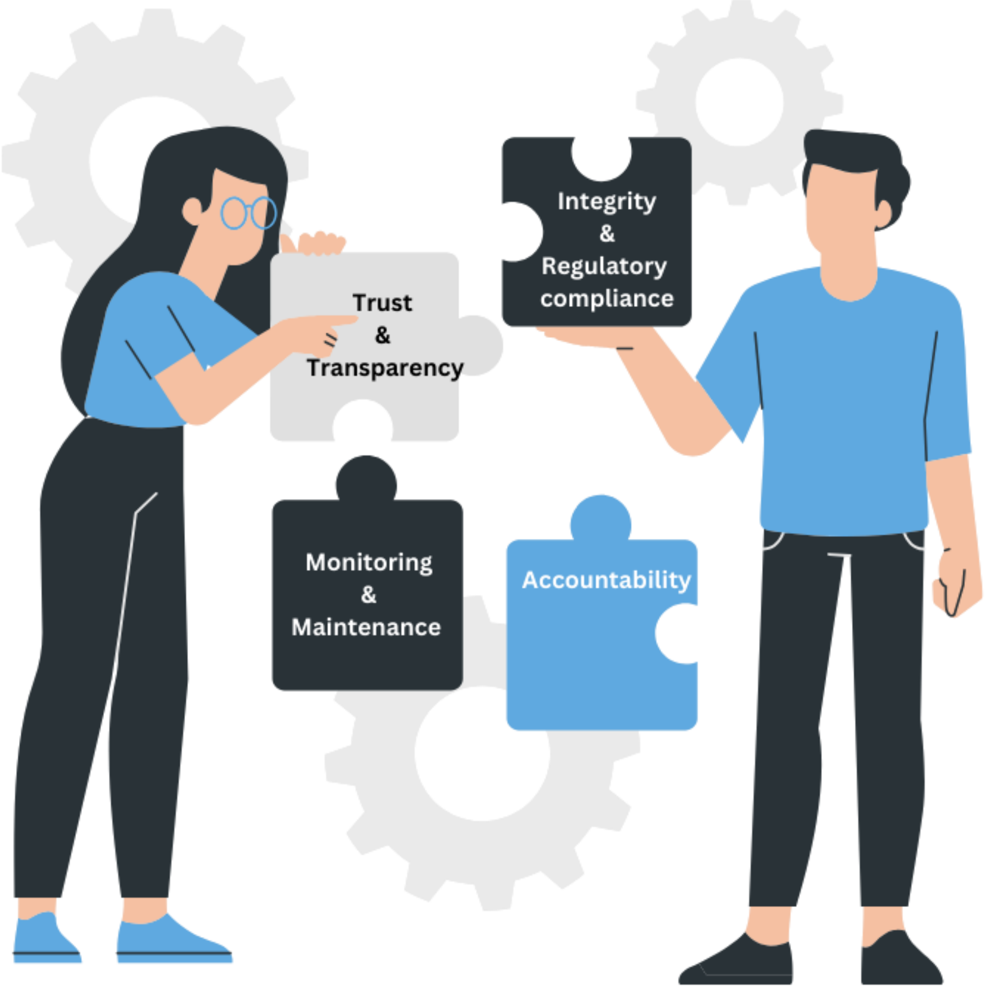
1. **Trust & Transparency**

   ❑ *Inaccessible or Non-Traceable Logs:* Logs that detail the model's process are not traceable or not accessible.

2. **Integrity & Regulatory compliance**

   ❑ *Claims are not verifiable:* For instance, the claimed accuracy of a diagnostic model might not be reproducible in external studies.

   ❑ A data provider shares a model card for a specific purpose, but it is repurposed for another use, potentially violating **GDPR** rules and patient consent.



Introduction (7/7)

Problematics

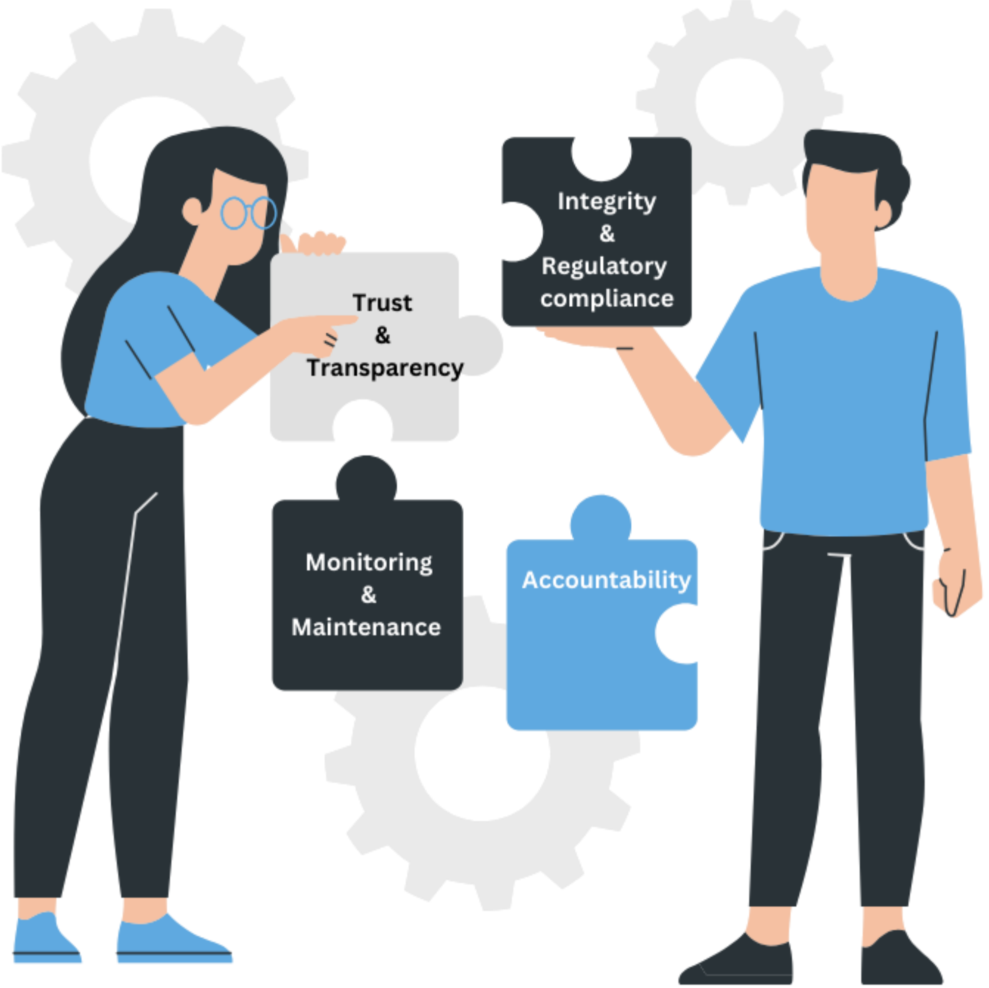
3. Monitoring & Maintenance

❑ *Metrics are not secure:* Performance metrics could be manipulated or not robustly validated.

❑ An outdated model continues to be used without re-evaluation, potentially leading to incorrect diagnoses due to new medical knowledge not being incorporated.

4. Accountability

❑ *Unclear responsibility*: It is not clear who is responsible; **e.g.,** in the event of a misdiagnosis, it remains ambiguous whether the model developers, the data providers, or another party should be held accountable.



Our Contribution (1/5)

Proposed Approach

- ❑ **Smart Wallet Contract (Managed Smart Wallet - EIP 4337):**
  - ▪ *Deployment of Adapted Managed Factory Contract*
  - ▪ Entity Registration and Role Management

Key Benefits

- *Enhanced User Autonomy and Security*: Features like transaction batching and sponsored transactions improve user experience in managing model cards.

- *Seamless Authentication and Upgradeability*: Provides a straightforward login, leading to secure user authentication and wallet upgradeability via smart contracts.

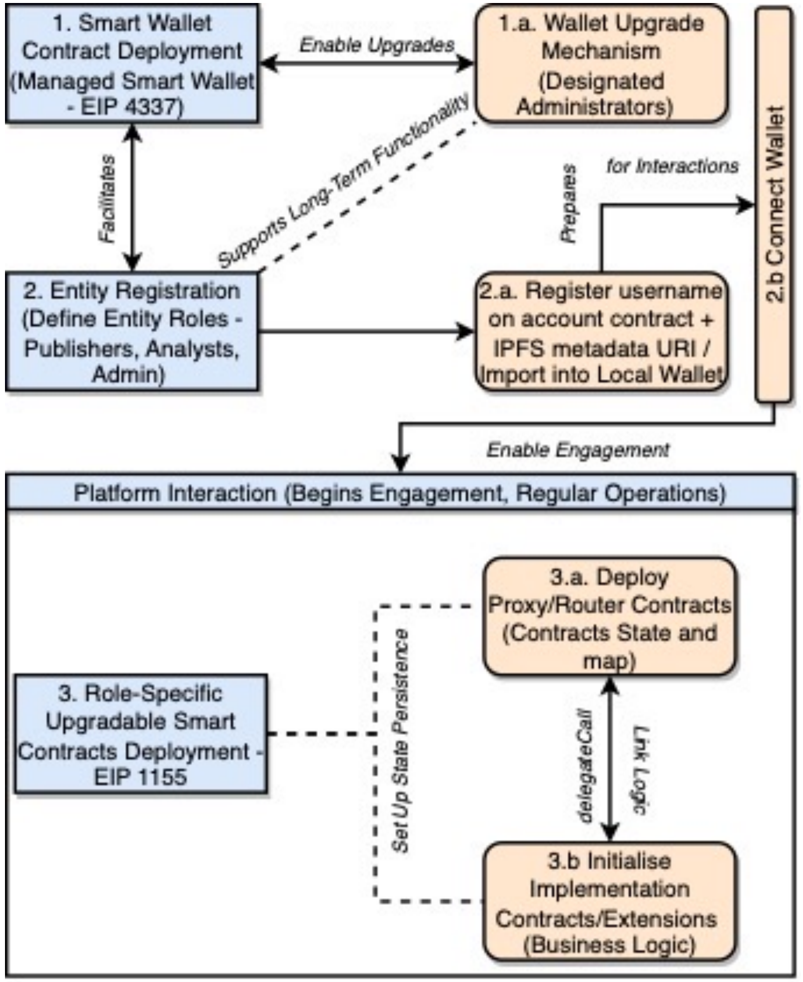

1. Smart Wallet Contract Deployment (Managed Smart Wallet - EIP 4337)
Enable Upgrades
1.a. Wallet Upgrade Mechanism (Designated Administrators)
Facilitates
Supports Long-Term Functionality
Prepares
for Interactions
2.b Connect Wallet
2. Entity Registration (Define Entity Roles - Publishers, Analysts, Admin)
2.a. Register username on account contract + IPFS metadata URI / Import into Local Wallet
Enable Engagement
Platform Interaction (Begins Engagement, Regular Operations)
3.a. Deploy Proxy/Router Contracts (Contracts State and map)
3. Role-Specific Upgradable Smart Contracts Deployment - EIP 1155
Set Up State Persistence
delegateCall
Link Logic
3.b Initialise Implementation Contracts/Extensions (Business Logic)

Figure 4: An Overview of the Proposed Approach (*m-LUCE*)

Our Contribution (2/5)

Proposed Approach

- □ **Role-Specific Upgradable Smart Contracts Deployment - EIP 1155**
  - Incorporates ERC-4907 extensions to set specific access durations for model cards, with automatic access revocation.

Key Benefits

- *Multi-Functionality with EIP 1155:* Supports issuing and distributing model cards to multiple analysts simultaneously, leveraging batch minting capabilities.

- *Upgradeability Through Proxy/Router Contracts: To support* system's flexibility via upgradeable contract structures.

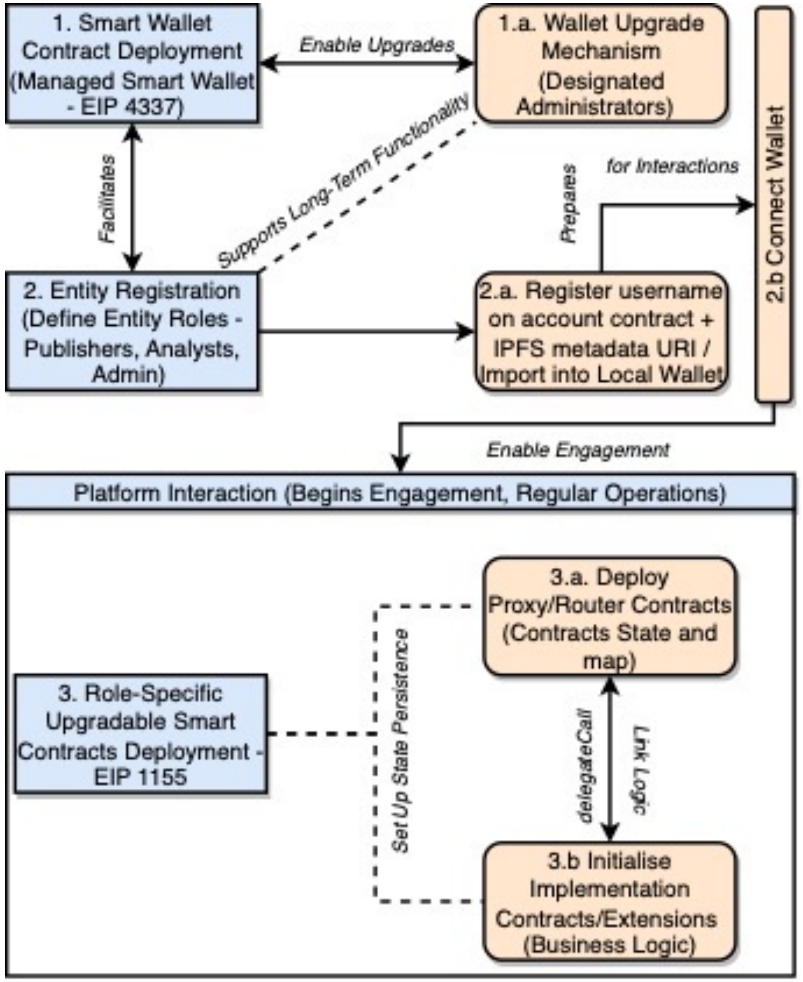

1. Smart Wallet Contract Deployment (Managed Smart Wallet - EIP 4337)
Enable Upgrades
1.a. Wallet Upgrade Mechanism (Designated Administrators)
Facilitates
Supports Long-Term Functionality
Prepares
for Interactions
2.b Connect Wallet
2. Entity Registration (Define Entity Roles - Publishers, Analysts, Admin)
2.a. Register username on account contract + IPFS metadata URI / Import into Local Wallet
Enable Engagement
Platform Interaction (Begins Engagement, Regular Operations)
3.a. Deploy Proxy/Router Contracts (Contracts State and map)
Set Up State Persistence
3. Role-Specific Upgradable Smart Contracts Deployment - EIP 1155
delegateCall
Link Logic
3.b Initialise Implementation Contracts/Extensions (Business Logic)

Figure 4 (Cont..): An Overview of the Proposed Approach (*m-LUCE*)

Our Contribution (3/5)

Roles and Interactions

1. ***Submit Claim***

   - ❑*Network Participant (**NP**), acting as the **Provider (a kind of an AI Developer)**, submits a claim to the m-LUCE.*

2. ***Process Claim***

   - ❑*The **NP**, serving as the **Evaluator (a kind of a certification authority)**, processes the claim, including submitting the claim for evaluation, requesting revisions, processing revisions, validating, and potentially rejecting the claim, all within the m-LUCE framework.*

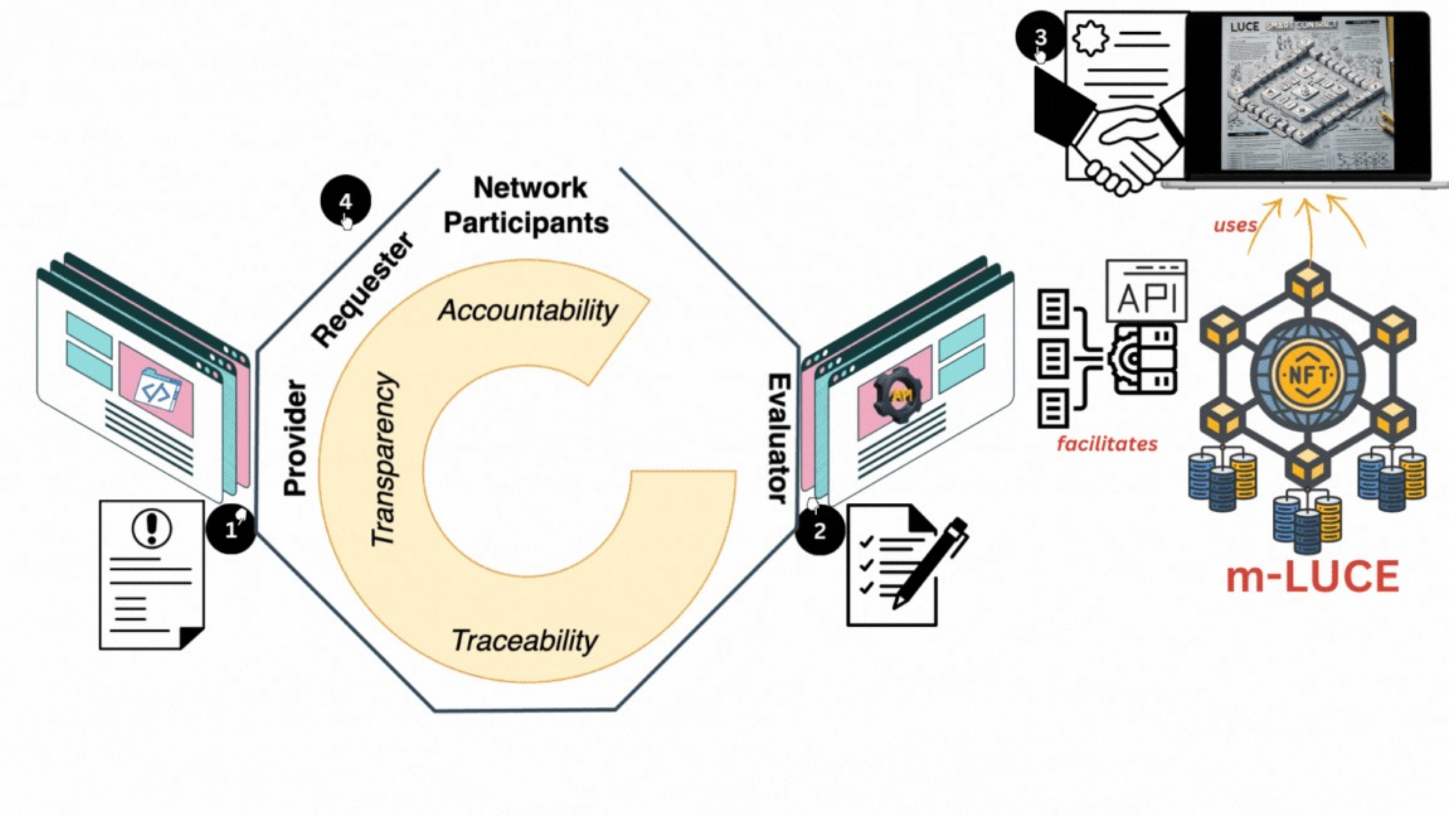


Figure 5: Big Picture of Network Participant's Journey in m-LUCE

Our Contribution (4/5)

Roles and Interactions

3. *Publish Claim*

- ❑*Once validated, the **NP** (**Evaluator**) may publish the claim.*

4. *Access Claim*

- ❑*An **NP** (**Requester**) searches for claims of interest. Access to these claims is contingent on permissions defined by m-LUCE's smart contract. Based on these permissions, access will either be granted or denied.*

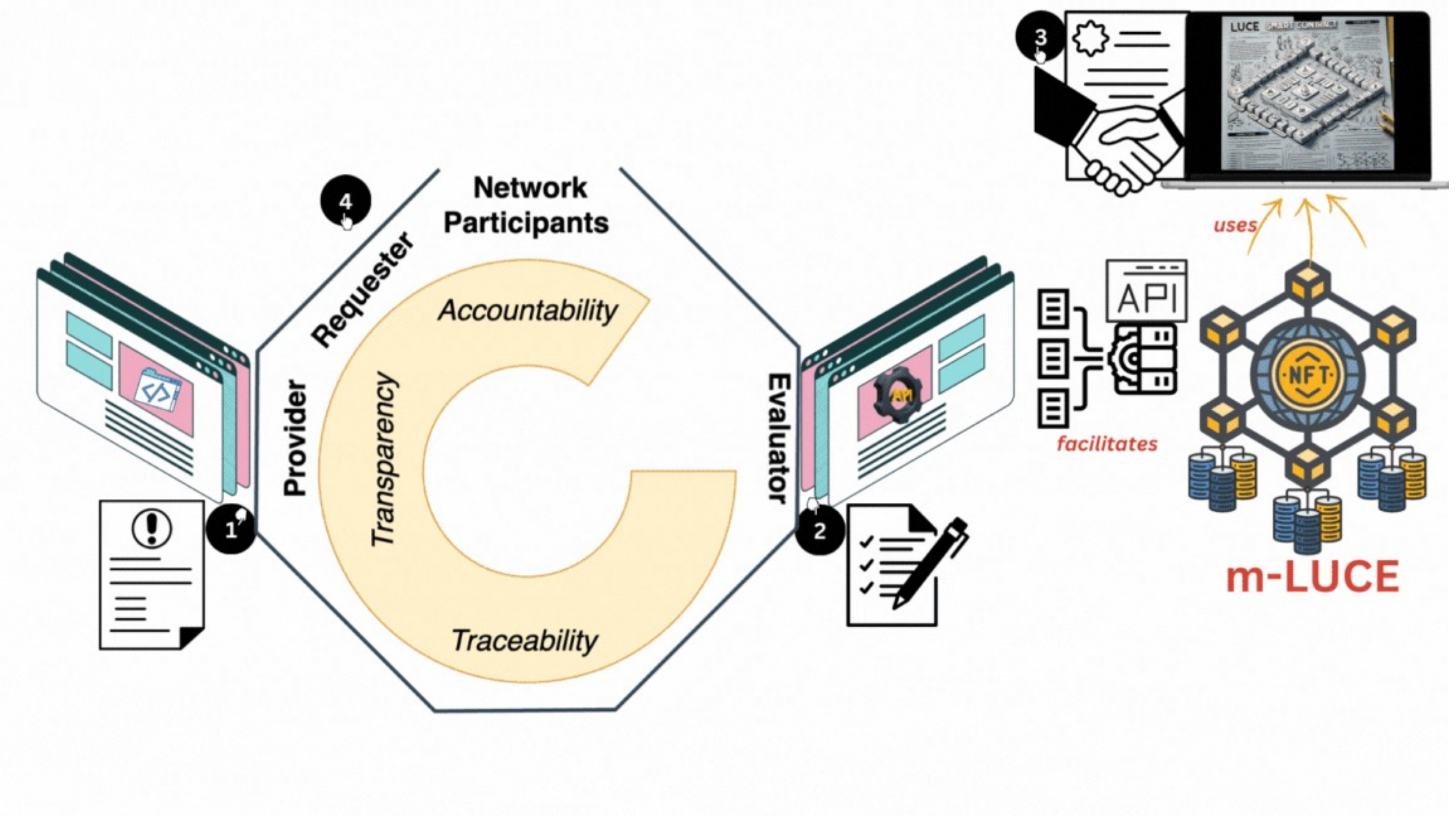
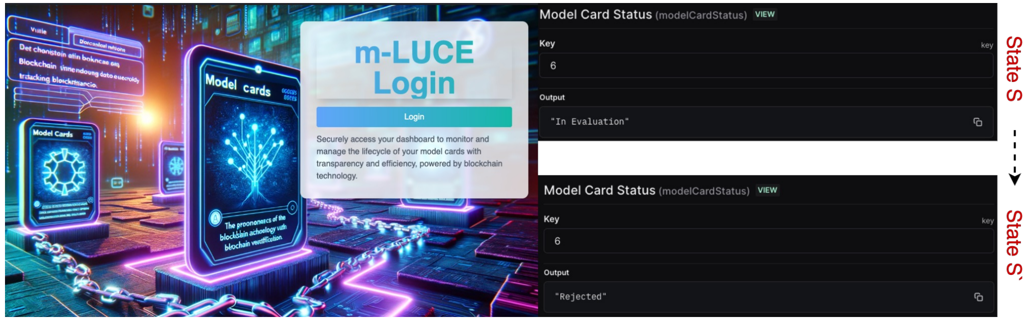


Figure 5 (Cont..): Big Picture of Network Participant's Journey in m-LUCE

Our Contribution (5/5)

Model Card Lifecycle

□ Developer



m-LUCE Login

Login

Securely access your dashboard to monitor and manage the lifecycle of your model cards with transparency and efficiency, powered by blockchain technology.

Model Card Status (modelCardStatus) VIEW

Key 6

Output "In Evaluation"

State S

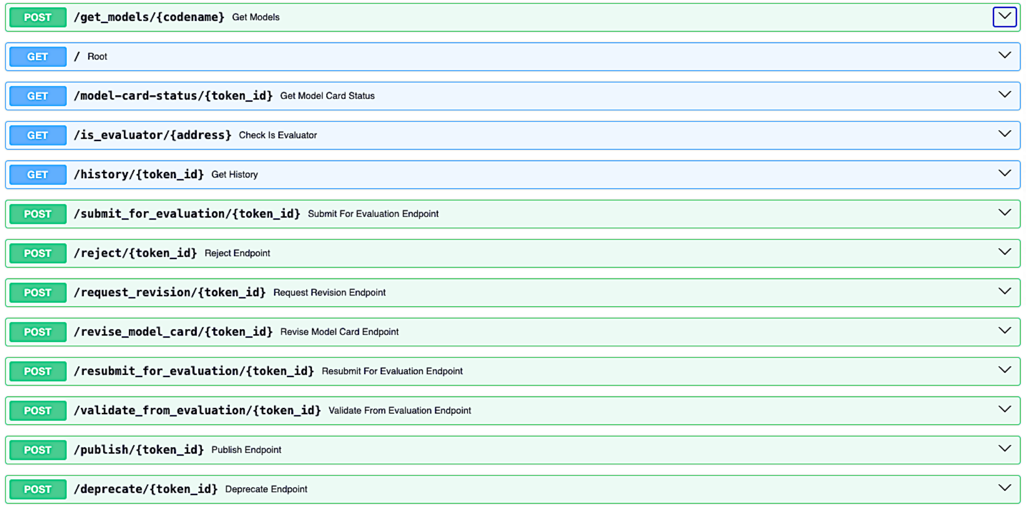
Model Card Status (modelCardStatus) VIEW

Key 6

Output "Rejected"

State S`

□ Evaluator & Requester



POST /get_models/{codename} Get Models

GET / Root

GET /model-card-status/{token_id} Get Model Card Status

GET /is_evaluator/{address} Check Is Evaluator

GET /history/{token_id} Get History

POST /submit_for_evaluation/{token_id} Submit For Evaluation Endpoint

POST /reject/{token_id} Reject Endpoint

POST /request_revision/{token_id} Request Revision Endpoint

POST /revise_model_card/{token_id} Revise Model Card Endpoint

POST /resubmit_for_evaluation/{token_id} Resubmit For Evaluation Endpoint

POST /validate_from_evaluation/{token_id} Validate From Evaluation Endpoint

POST /publish/{token_id} Publish Endpoint

POST /deprecate/{token_id} Deprecate Endpoint

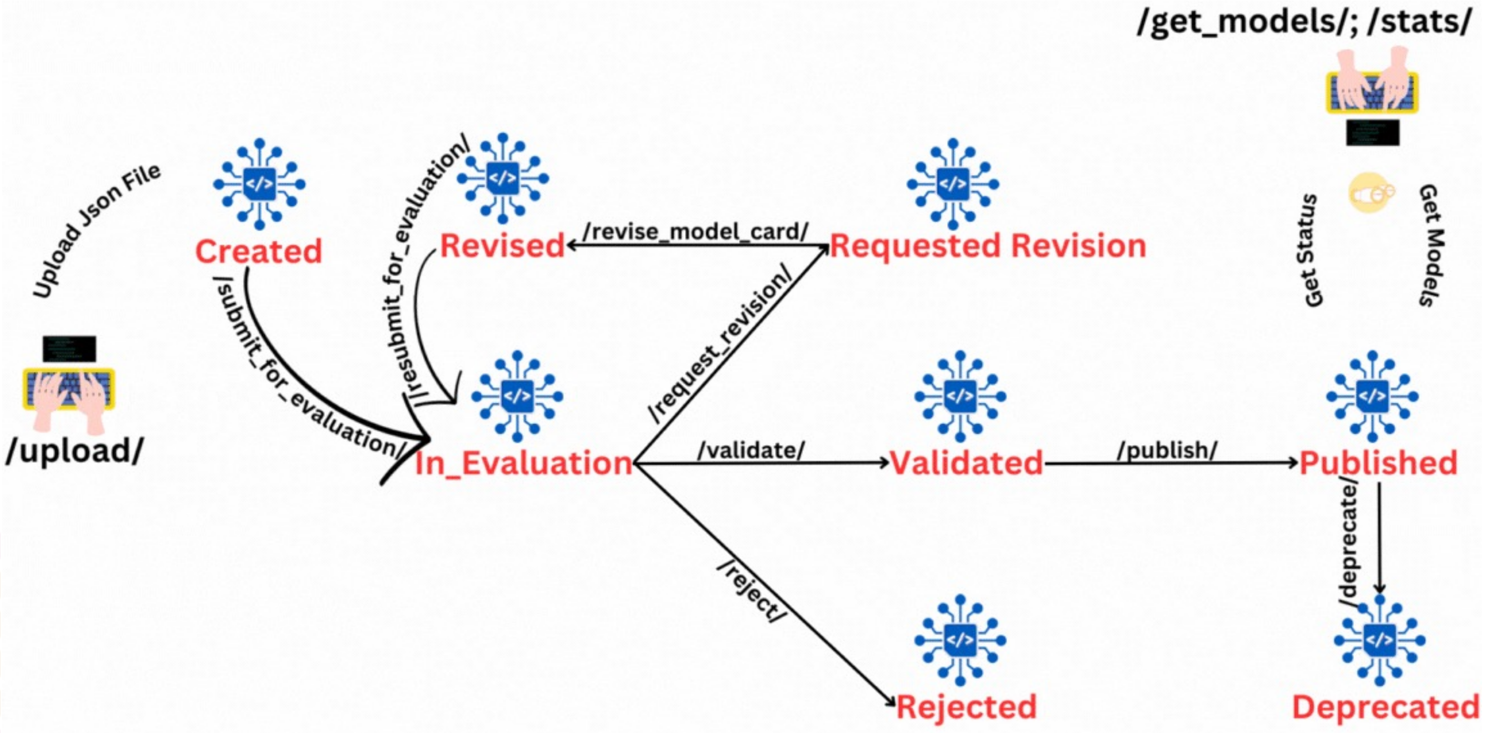


Figure 6: Illustrating Possible States and Transitions of Model Cards

Some of the Key Takeaways (1/3)

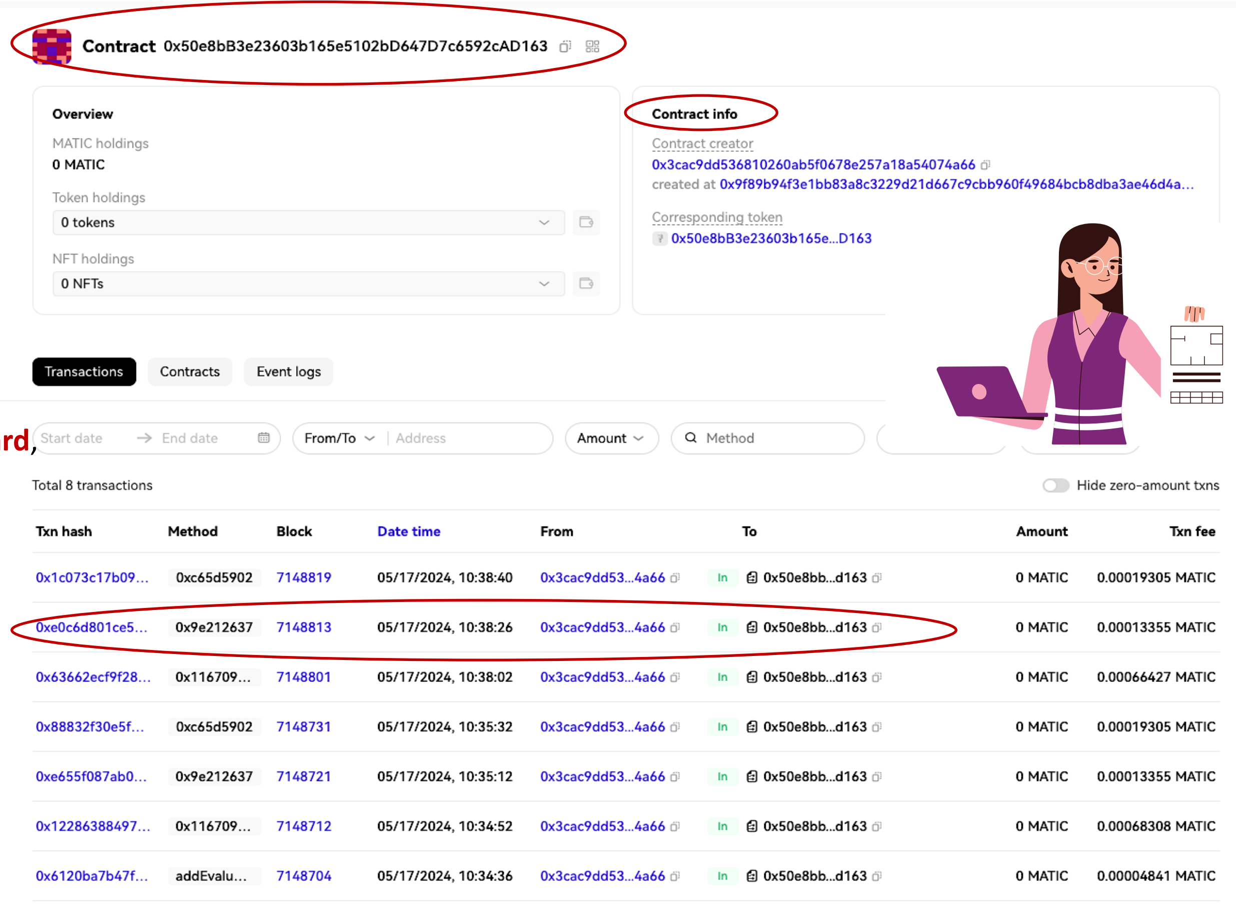
Transparency

❑ ***How can logging every transaction and contract interaction in real-time enhance the transparency, and integrity, of the model card's lifecycle?***

1. **Public Accessibility of Data** (e.g., *Smart Contracts* and *Transactions*)
   - a) ***Contract address:*** 0x50e8bB3e23603b165e5102bD647D7c6592cAD163
   - b) ***Full transaction history and details*** (e.g., *Txn hash*, *Date time*, *From*, *To*

2. **Enhanced Trust**
   - a) *Creator Information:* 0x3cac9dd536810260ab5f0678e257a18a54074a66
   - ■ *Outcome:* Stakeholders can verify the authenticity and origin of the **Model Card**, knowing exactly who created it

3. **Pseudonymity**
   - a) *Maintaining Privacy with Transparency*
   - ■ *Outcome: Instead of using personal information like **usernames** (e.g., Ankur), blockchain uses **public addresses** to maintain user privacy, striking a balance between transparency and privacy, which fosters trust and security*



Contract 0x50e8bB3e23603b165e5102bD647D7c6592cAD163

Overview

MATIC holdings
0 MATIC

Token holdings
0 tokens

NFT holdings
0 NFTs

Contract info

Contract creator
0x3cac9dd536810260ab5f0678e257a18a54074a66
created at 0x9f89b94f3e1bb83a8c3229d21d667c9cbb960f49684bcb8dba3ae46d4a...

Corresponding token
0x50e8bB3e23603b165e...D163

Total 8 transactions

Hide zero-amount txns

Txn hash	Method	Block	Date time	From		To	Amount	Txn fee
0x1c073c17b09...	0xc65d5902	7148819	05/17/2024, 10:38:40	0x3cac9dd53...4a66	In	0x50e8bb...d163	0 MATIC	0.00019305 MATIC
0xe0c6d801ce5...	0x9e212637	7148813	05/17/2024, 10:38:26	0x3cac9dd53...4a66	In	0x50e8bb...d163	0 MATIC	0.00013355 MATIC
0x63662ecf9f28...	0x116709...	7148801	05/17/2024, 10:38:02	0x3cac9dd53...4a66	In	0x50e8bb...d163	0 MATIC	0.00066427 MATIC
0x88832f30e5f...	0xc65d5902	7148731	05/17/2024, 10:35:32	0x3cac9dd53...4a66	In	0x50e8bb...d163	0 MATIC	0.00019305 MATIC
0xe655f087ab0...	0x9e212637	7148721	05/17/2024, 10:35:12	0x3cac9dd53...4a66	In	0x50e8bb...d163	0 MATIC	0.00013355 MATIC
0x12286388497...	0x116709...	7148712	05/17/2024, 10:34:52	0x3cac9dd53...4a66	In	0x50e8bb...d163	0 MATIC	0.00068308 MATIC
0x6120ba7b47f...	addEvalu...	7148704	05/17/2024, 10:34:36	0x3cac9dd53...4a66	In	0x50e8bb...d163	0 MATIC	0.00004841 MATIC

Figure 7: A Snippet of a Transaction History

Some of the Key Takeaways (2/3)

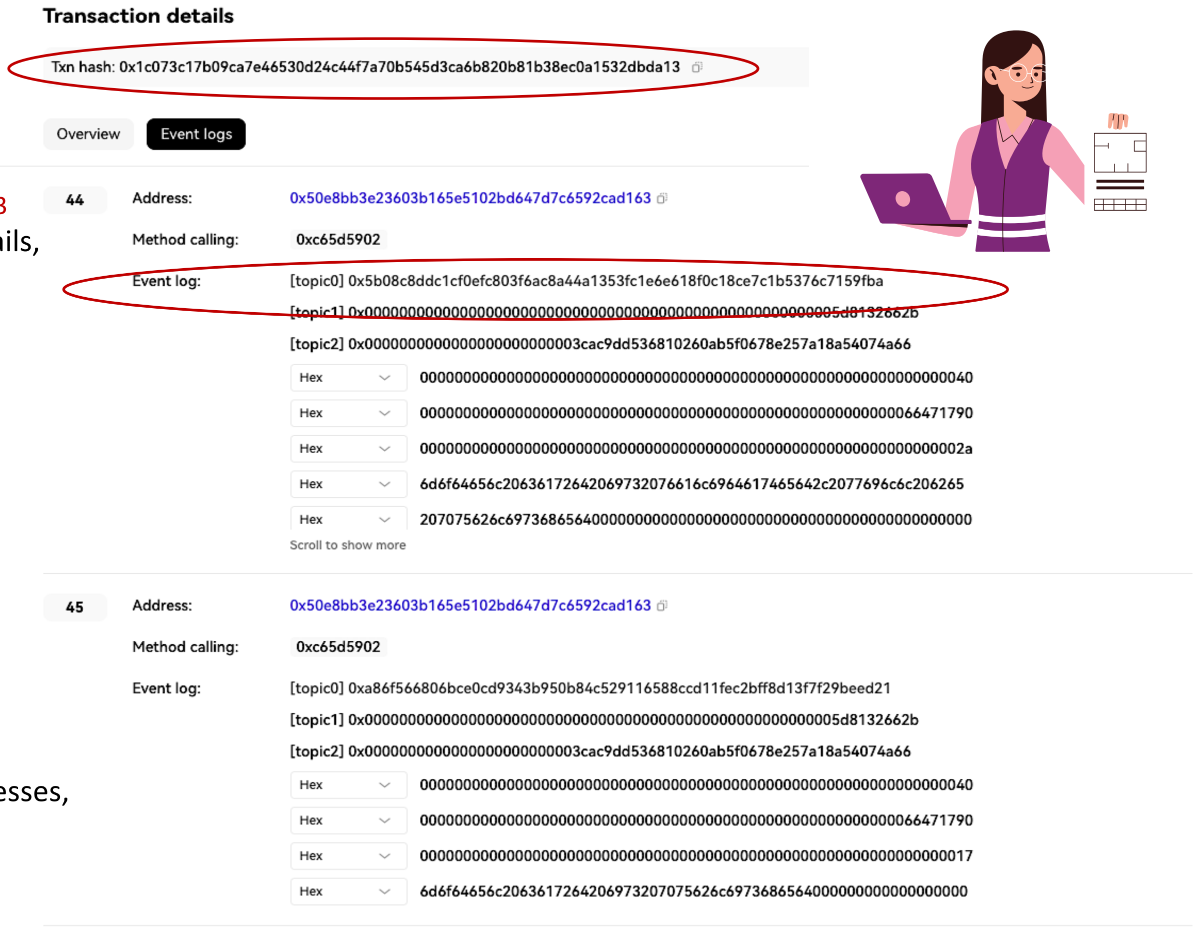
Traceability

❑ ***How does tracking every step in the lifecycle of a Model Card, from creation to interaction, ensure a comprehensive record of all activities?***

1. **Initiate a Transaction Lookup**
   - *a)* ***Transaction hash***: 0x1c073c17b09ca7e46530d24c44f7a70b545d3c6ab820b81b38ec0a1532dbda13
   - ■ *Outcome:* Uniquely identifies the transaction, allowing anyone to trace and verify all details, including the involved addresses and method called in the lifecycle of a **Model Card**

2. **Review Event Logs** (e.g., *What occurred during the transaction*)
   - *a) Methods were called on the contract and the data associated with these calls*
   - ■ *Method Calling (0xc65d5902)* -- *e.g., transfer(address to, uint256 value)*
   - ■ *Event Log:*
     - ○ *topic 0: Event signature (e.g., Transfer(address indexed from, address indexed to, uint256 value))*
     - ○ *topic 1: First indexed parameter (e.g., sender address)*
     - ○ *topic 2: Second indexed parameter (e.g., receiver address)*

3. **Trace the Participants, Verify and Cross-reference**
   - *a) The addresses involved (From and To) are clearly recorded, making it possible to trace that address 0x3cac9dd536810260ab5f0678e257a18a54074a66 initiated the interaction with the contract at 0x50e8bB3e23603b165e5102bD647D7c6592cAD163.*
   - ■ *Outcome:* By cross-referencing multiple transactions and event logs involving these addresses, you can build a comprehensive picture of activities and interactions related to model cards.



Transaction details

Txn hash: 0x1c073c17b09ca7e46530d24c44f7a70b545d3ca6b820b81b38ec0a1532dbda13

44 Address: 0x50e8bb3e23603b165e5102bd647d7c6592cad163

Method calling: 0xc65d5902

Event log:

```
[topic0] 0x5b08c8ddc1cf0efc803f6ac8a44a1353fc1e6e618f0c18ce7c1b5376c7159fba
[topic1] 0x0000000000000000000000000000000000000000000000000000005d8132662b
[topic2] 0x0000000000000000000000003cac9dd536810260ab5f0678e257a18a54074a66
Hex  0000000000000000000000000000000000000000000000000000000000000040
Hex  0000000000000000000000000000000000000000000000000000000066471790
Hex  000000000000000000000000000000000000000000000000000000000000002a
Hex  6d6f64656c2063617264206973207661696461746564 2c2077696c6c206265
Hex  207075626c69736865640000000000000000000000000000000000000000000000
```

Scroll to show more

45 Address: 0x50e8bb3e23603b165e5102bd647d7c6592cad163

Method calling: 0xc65d5902

Event log:

```
[topic0] 0xa86f566806bce0cd9343b950b84c529116588ccd11fec2bff8d13f7f29beed21
[topic1] 0x0000000000000000000000000000000000000000000000000000005d8132662b
[topic2] 0x0000000000000000000000003cac9dd536810260ab5f0678e257a18a54074a66
Hex  0000000000000000000000000000000000000000000000000000000000000040
Hex  0000000000000000000000000000000000000000000000000000000066471790
Hex  0000000000000000000000000000000000000000000000000000000000000017
Hex  6d6f64656c20636172642069732070756 26c69736865640000000000000000000
```

Figure 8: A Snippet of Transaction Event Logs

Some of the Key Takeaways (3/3)

Accountability

❑ ***How does ensuring accountability for all participants throughout the lifecycle of a model card enhance the integrity and security of the entire process?***

1. **Ownership Tracking (e.g., current holder of the token)**

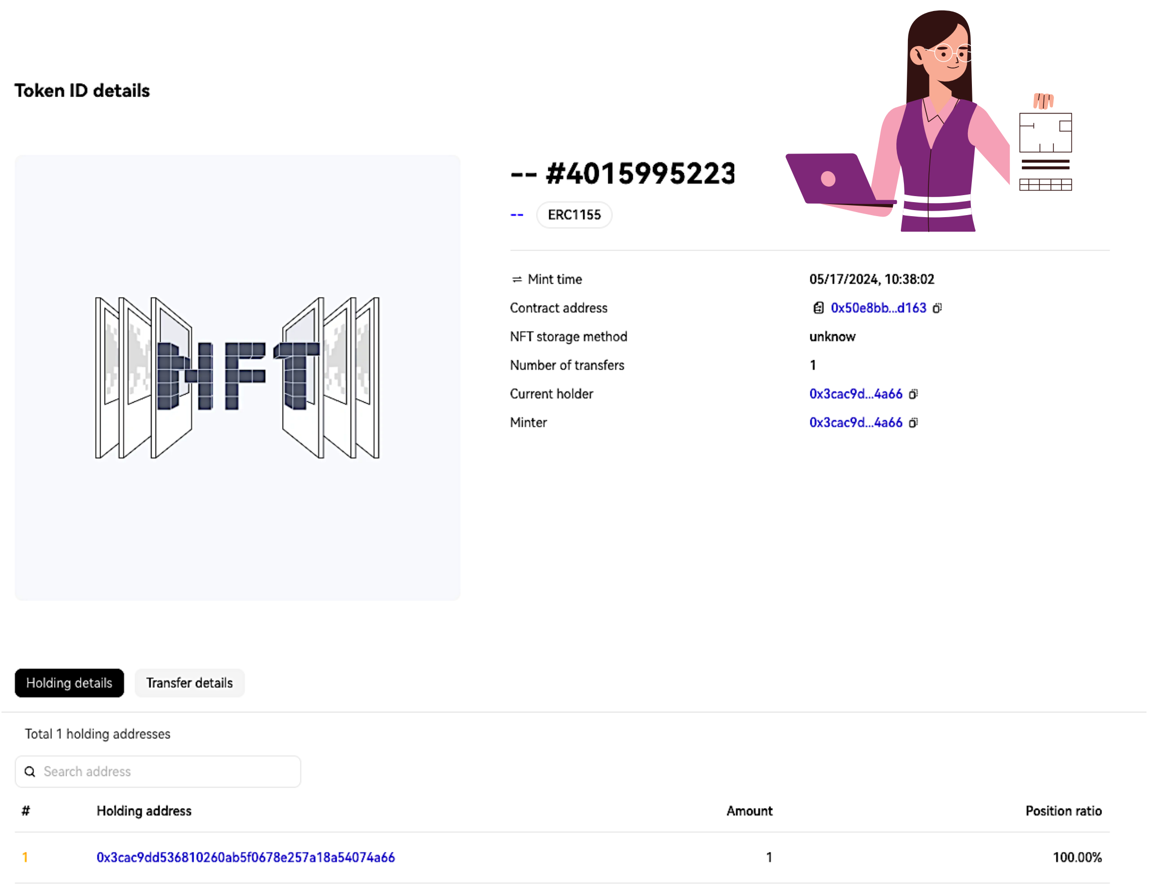
   a) Token ID Details: e.g., #401599522347

   ■ *Outcome:* Uniquely identifies the transaction, allowing anyone to trace and verify all details, including the involved addresses and method called in the lifecycle of a **Model Card**

2. **Immutable Logs of Action**

   a) Once the model card is minted and any transactions are recorded, these records are permanently recorded, making them impossible to alter or delete.

   ■ *Outcome:* Ensures that the historical data of the Model Card is protected from tampering, thereby enhancing accountability and trust



Token ID details

-- #4015995223

-- ERC1155

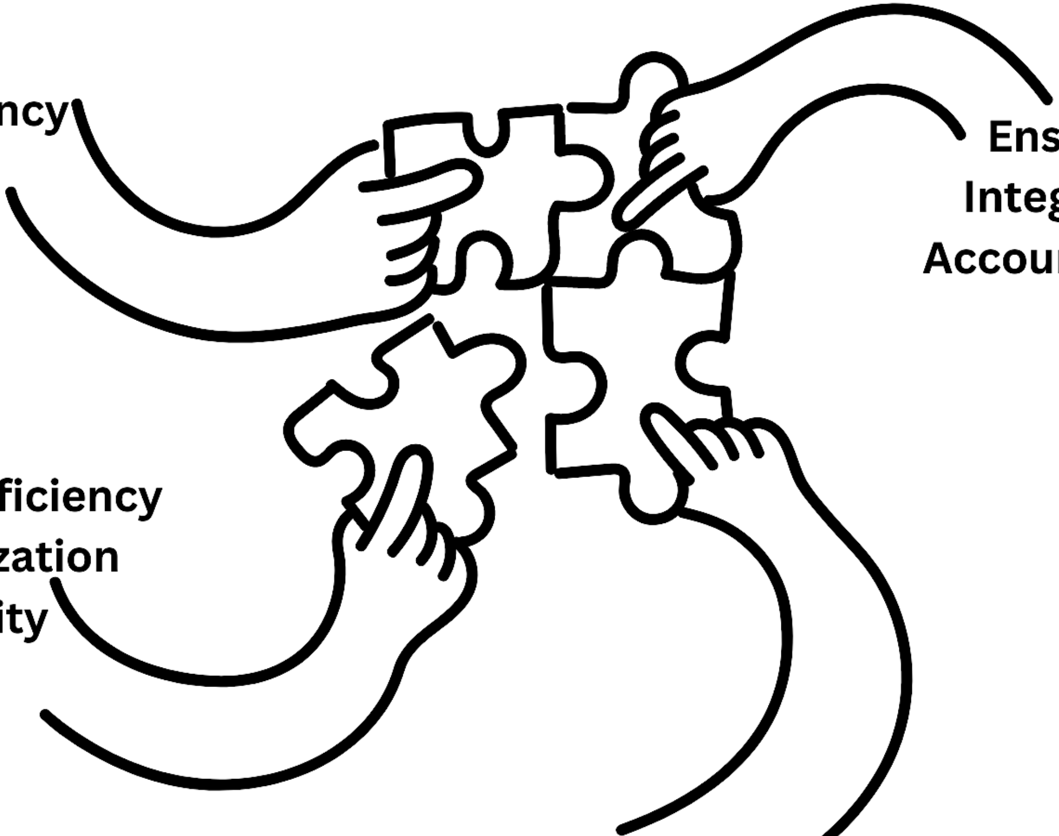
⇌ Mint time	05/17/2024, 10:38:02
Contract address	0x50e8bb...d163
NFT storage method	unknow
Number of transfers	1
Current holder	0x3cac9d...4a66
Minter	0x3cac9d...4a66

Total 1 holding addresses

#	Holding address	Amount	Position ratio
1	0x3cac9dd536810260ab5f0678e257a18a54074a66	1	100.00%

Figure 9: A Snippet of a Created Model Card as an NFT using ERC-1155

Conclusion and Perspectives



Enhanced Transparency & Efficacy

Ensured Integrity & Accountability

- Exploring System Efficiency
- Expanding Tokenization and Applicability

Robust Lifecycle Management

References

1. Perrault, R., & Clark, J. (2024). Artificial Intelligence Index Report 2024. Retrieved from https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_2024_AI-Index-Report.pdf
2. Liang, Weixin, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. "What's documented in AI? Systematic Analysis of 32K AI Model Cards." arXiv preprint arXiv:2402.05160 (2024).
3. Google. About Model Cards. Retrieved from https://modelcards.withgoogle.com/about
4. Hugging Face. Model Cards. Retrieved from https://huggingface.co/docs/hub/en/model-cards
5. European Parliament. (2024). Artificial Intelligence Act. Retrieved from https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

Thank you for your attention!

Any Questions

Ankur Lohachab (Postdoctoral Researcher)

ankur.lohachab@maastrichtuniversity.nl; ankurlohachab@gmail.com

Institute of Data Science, Department of Advanced Computing Sciences, Maastricht University, Maastricht, Netherlands