

# AI/ML Model Cards in Edge AI Cyberinfrastructure: towards Agentic AI

Beth Plale

School of Computer and Data Sciences  
University of Oregon  
Eugene, OR USA  
0000-0003-2164-8132

Neelesh Karthikeyan

School of Computer and Data Sciences  
University of Oregon  
Eugene, OR USA  
0009-0009-9817-7042

Isuru Gamage

Intelligent Systems Engineering  
Indiana University  
Bloomington, IN USA  
agamage@iu.edu

Joe Stubbs

Texas Advanced Computing Center (TACC)  
University of Texas  
Austin, TX USA  
jstubbs@tacc.utexas.edu

Sachith Withana

School of Computer and Data Sciences  
University of Oregon  
Eugene, OR USA  
0000-0002-2022-8155

**Abstract**—AI/ML model cards can contain a benchmarked evaluation of an AI/ML model against intended use but a one time assessment during model training does not get at how and where a model is actually used over its lifetime. Through Patra Model Cards embedded in the ICICLE AI Institute software ecosystem we study model cards as dynamic objects. The study reported here assesses the benefits and tradeoffs of adopting the Model Context Protocol (MCP) as an interface to the Patra Model Card server. Quantitative assessment shows the overhead of MCP as compared to a REST interface. The core question however is of active sessions enabled by MCP; this is a qualitative question of fit and use in the context of dynamic model cards that we address as well.

**Index Terms**—AI accountability, Model Context Protocol(MCP), AI/ML Model Cards,

## I. INTRODUCTION

AI/ML models are widely used in scientific research. Their traceability (where and in what manner used, how one AI/ML model is related to another, etc) is often overlooked. But for responsible use of AI and to increase the reproducibility of science, this traceability is important. AI/ML model cards are a step in the right direction.

An AI/ML *Model Card* is a structured overview of how an AI/ML model was designed and evaluated. Mitchell et al. describe the original intent behind model cards this way: "model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic subgroups that are relevant to the intended application domains" [9]. Model cards further are used to disclose the context under which models are intended to be used, details of the performance evaluation procedures, and other relevant information. As Mitchel et al. state, model

cards provide benchmarked evaluation that is necessary for the informed use of a model.

Model Cards, which have seen wide adoption since their introduction, are an entry into strengthening *responsible AI practice* which itself includes *explainability* which provides humans with clear, understandable, and meaningful explanations for why an AI system made a certain decision, *interpretability*, which focuses on understanding the inner workings of how a model's inputs influence its outputs, and *accountability* which gets at who is accountable for the AI systems when something goes wrong. Big tech companies like Amazon, Hugging Face, and Google are prominent in their support for model cards and provide tools for card creation and persistence. For instance, the Amazon SageMaker has its Model Card that is integrated in the SageMaker Model Registry. With the broad uptake of Large Language Models, the debate around responsible AI has intensified. And some tech providers are responding. For instance, in summer of 2023 Meta and Microsoft announced the launch of Llama2. The cited research paper behind the release is extensive in its examination of the model through the lense of responsible AI [20]. The paper includes its Model Card in the appendix (though the Model Card hews to the original Mitchell et al intent and uses the research paper itself for greater exposition.)

We acknowledge and celebrate all forms of AI/ML model cards as a step in the right direction. *Our broader vision of the utility of AI/ML model cards focuses on whether benchmarked evaluation of an AI/ML model can be treated not as a one-time thing, done during model training, but instead can be used to assess and evolve one's understanding of a model's fit for purpose during its lifetime of use.* Can we particularly collect information that might point to an ill suited use? This problem has motivated us to develop the Patra Model Card framework [24]. We view an AI/ML model card not just as manual documentation created at time of model creation, but as an active reflection of model usage through time which can,

thus, if practicable, make greater contributions to responsible AI. Our study is focused on AI/ML inference at the edge, and is carried out in the context of the evolving ICICLE AI Institute computational environment [10].

Edge environments present unique challenges (resource constraints, workload variability, and discrete changes in quality requirements) that static model deployment strategies cannot adequately address. For instance, camera traps in wildlife monitoring may experience sudden bursts when animals appear, requiring rapid adaptation between accuracy-optimized and latency-optimized models. Traditional edge AI deployments rely on fixed model choices that either optimize for performance or require manual reconfiguration when conditions change. This approach fails to capture the dynamic nature of edge workloads, where motion detection, lighting changes, or seasonal patterns can dramatically shift the optimal balance between inference accuracy and response time.

We have gained experience with model cards through Patra’s use in the NSF AI Intelligent CI with Computational Learning in the Environment (ICICLE) Institute [10] which is developing novel AI-driven pipelines for scientific research at the edge. This is through, in part, by working with MLFieldPlanner, a configurable cyberinfrastructure that can utilize the full edge-cloud-HPC continuum, analyze ML pipelines, and study edge-to-center tradeoffs in model placement. Researchers use ML Field Planner to “configure experiments to run on real IoT hardware, configure machine learning models to analyze custom benchmark datasets, and experiment with different algorithm configurations, such as storage compression, all from a graphical user interface” [17].

The motivating use case for our work is AI/ML model deployment at the edge in support of field research [17]. The scenario is of a researcher deploying AI-enabled devices for multi-modal observation in a remote environmental setting. For example, an ecologist may evaluate the compatibility, latency, and recall of different versions of a YOLO object detection model [14] deployed to various edge devices, such as Raspberry Pis and the NVIDIA Jetson family, each equipped with a motion-activated camera. Patra Model Cards coordinate and interoperate with the TapisUI workflow platform [17] to facilitate edge-based, plug-and-play model orchestration. Through TapisUI, a scientist queries Patra for models that fit their needs (for example, models optimized for less than 100 ms edge inference latency). Relevant candidate models are directly retrieved from their distributed hosting platforms such as HuggingFace or GitHub through persistent IDs maintained in the model card.

As we have accumulated experience with use of model cards in ICICLE, and accumulated the model cards themselves, we undertook to explore how the Patra Model Card repository can beneficially serve in an agentic AI setting. An Agentic AI system is an AI system that exhibits adaptability in achieving complex goals in complex environments with limited direct supervision [25]. We particularly study the tradeoffs of supporting the Model Context Protocol (MCP) [8], an interoperability protocol for connecting AI applications to

external systems. MCP offers itself as a standardized way for AI applications to connect to external systems for “connecting AI assistants to the systems where data lives, including content repositories, business tools, and development environments” [8].

We make the following observations about the Patra framework as we consider the viability of MCP as a replacement or additional interface to the REST API. These observations guide and constrain next steps:

- 1) When models are frequently deployed their model cards are subject to continuous streaming ingest of information about model use in the ICICLE environment. A model card thus become quite large (in the GB range.)
- 2) Patra uses a graph database (currently Neo4j and Cypher queries) that is fronted with a rich API implemented as Flask endpoints. Graph databases have rudimentary schemas, that is, of edges and vertices (nodes). Patra layers meaning on top of this rudimentary schema through 1) ascribing properties to nodes (and making heavy use of these properties) and 2) through embedding pathways and delineations of subgraphs by logic in the endpoints.

In this article and contextualized within the larger setting of the ICICLE Institute software ecosystem, we make the following contributions:

- foundational principles behind Model Cards as dynamic objects
- use case of model cards as dynamic objects in their integration and use in the ICICLE ecosystem
- quantitative results of two MCP implementations in comparison to the currently supported REST API
- assessment of the benefits and tradeoffs of adopting the Model Context Protocol (MCP) as an interface to the Patra Model Card server. The core question is of active sessions enabled by MCP; this is a qualitative question of fit and use in the context of dynamic model cards.

The remainder of this article identifies the principles underlying the dynamism in model cards in Section II. Following that is a motivating use case in Section III, a discussion of system design and MCP in Section IV, experimental results in Section V, and a discussion in Section VI. Related work and Conclusions round the paper out in Sections VII and VIII respectively.

## II. METHODOLOGY

An AI/ML model card should tightly relate to the AI/ML model it describes - always accessible to lend clarity on what the model is, how it has been and will continue to be used. This is a tall order in an environment where AI/ML Models can be arbitrarily downloaded, transformed, and run anywhere. We have the benefit of exploring the utility of model cards by constraining the working environment to the broader ICICLE cyberinfrastructure environment. This execution environment is underpinned by the Tapis hosted API platform [18].

In this section we provide the foundation underlying Patra’s dynamic model card framework.

**AI/ML Model Lifecycle** We model edge AI/ML models deployed at the edge as being in one of the following four states (see Figure 1):

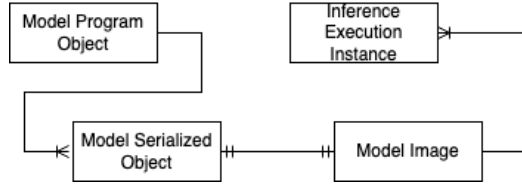


Fig. 1. AI/ML model object lifecycle

- **Model Program Object.** An AI/ML model begins its life in some state of training. Model Program Objects are often Python objects, are subject to training, and reside in a repository such as Github.
- **Model Serialized Object** has undergone serialization for sharing, packaging into a container, etc. Serialized models are persisted in an object store, file system, etc. Model serialization can occur any number of times so the n-ary relationship between a model program object and a model serialized object is 1:M.
- **Model Image** is a serialized model packaged for deployment as an inference server. This packaging makes a serialized model remotely accessible and deployable. A Model Image is frequently maintained as a persistent artifact in an image server such as DockerHub.
- **Inference Execution Instance** represents an inference server, a running executable. It is a time-bounded instance of model while executing. A single Model Image can be deployed repeatedly, hence a model image and its execution instance are in a 1:M relationship. While all artifacts have properties, the Inference Execution Instance artifact has temporal and spatial properties that capture its time and location of execution as well as such things as pattern of use.

**Model Development/Deployment Pipeline.** Patra Model Cards operate within a development pipeline in the ICICLE AI ecosystem. The pipeline has multiple phases from design through training, deployment, use, and analysis. This is depicted in Figure 2 starting in the upper right hand corner.

**Model Card as Design Aid.** The lifecycle of a deployment pipeline begins with a researcher/educator querying the Patra data source to gain an understanding of AI/ML models that are available for use, for instance by asking questions that get at issues of their prior history and use, such as: "What devices were used last time I ran this experiment?", "Who owns the computers that my model will be deployed to?", and "Tell me about the data used in last week's experiment. This activity is shown in the upper right bubble "Use Case Understanding".

**Model Onboarding.** The cross-hatch arrows pointing inward in the figure are static information collection points. During model training, a researcher interacts through their Jupyter notebook to provide basic information about their model (through the Patra ModelCard toolkit). The toolkit supports

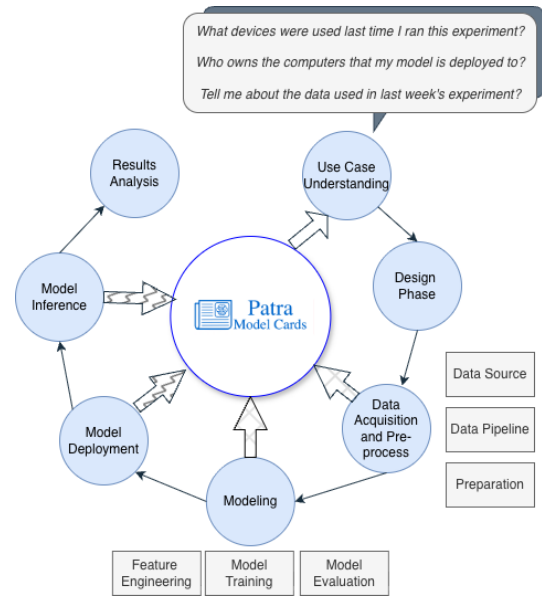


Fig. 2. AI/ML model development-deployment pipeline with model card collection points

auto-population of key metadata, for instance fairness and XAI scanners. The toolkit validates and outputs the Model Card as a structured JSON file. The actual model file (for example, a PyTorch .pt file) is uploaded to an artifact store such as HuggingFace or GitHub. The model card entry is updated with a link to the artifact's location. At this point, the model becomes available in the system's model catalog.

**Runtime Data Capture.** The zig-zag arrows pointing inward in Figure 2 are runtime information collection points. Patra continuously captures model behavior data through the CKN streaming system [23], our earlier work, during runtime inference at the edge. Each edge server runs a CKN daemon that streams events (prediction outcomes, accuracy, latency, CPU/GPU usage, etc.) through a Kafka message broker. These events are consumed by CKN stream processors; Patra subscribes to these streams which are the source of information about execution instances for the Model Card. The daemon captures real-time events from edge devices, augmenting them with details such as model usage, resource consumption, prediction accuracy, and latency. Patra then merges this data into its graph. Each inference execution becomes a node (an "execution instance" of the model) that is connected to the corresponding Model Card.

**Automated Model Selection.** While not directly reflected in Figure 2, when it is a system rather than a human choosing an inference model and because the entire Model Card is machine-actionable, an orchestrator can issue full-text or graph queries against Patra. In practice, a controller might call the /search endpoint with keywords or criteria (for example, "camera-trap classification") to retrieve matching Model Cards. It can filter results by fairness or explainability annotations embedded in the card. Alternatively, direct graph queries on performance attributes (e.g., selecting the model

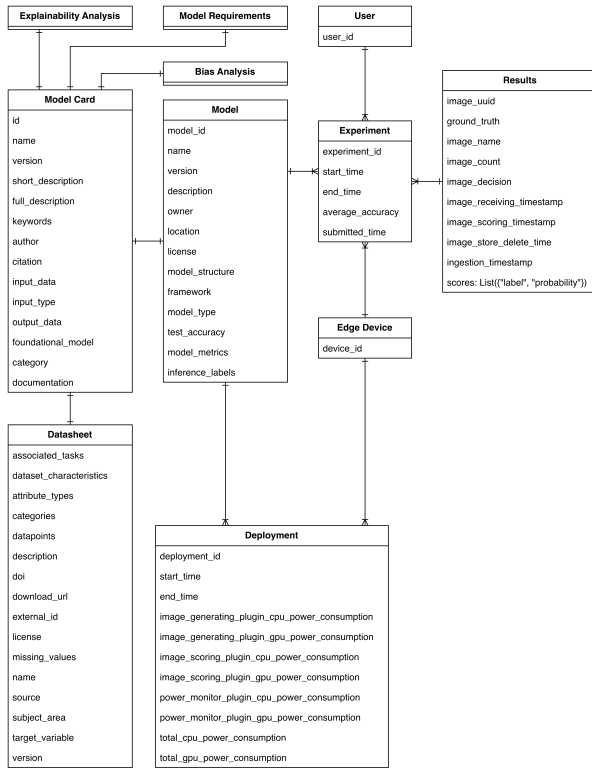


Fig. 3. Core Entities of the Patra Model Card

with the highest accuracy under a latency threshold) can be executed via Patra’s REST interface.

Patra provides a “machine-actionable API” enabling automated selection based on fairness, explainability, and performance metrics. In other words, the knowledge graph can be queried to rank models by these attributes. Once a suitable model ID is identified, its linkset can be resolved and the model artifact retrieved for deployment. These selection steps are grounded in the provenance recorded in Patra, so as to ensure that choices are traceable. If multiple models satisfy the criteria, Patra’s built-in similarity analysis can even infer versioning relations (e.g., `alternateOf` or `revisionOf`) to guide the selection of the most appropriate variant. Thus, automated model selection is achieved by programmatically querying the Patra-KG and using the integrated metadata and metrics to inform the decision.

**Core Entities.** The high level structure of a model card is captured in the E-R diagram in Figure 3. Shown are four nodes: Model Card, Model, Edge Server, and Deployment. Each node has properties (what is shown is a subset). Note that a Model Card node and Model node are in a 1:1 n-ary relationship. The model is described in its own node. A Model is in a relation with its Deployment. The relation (edge) is named, though the named edge is not depicted here. A model can be deployed repeatedly, as is captured by the 1:m n-ary relationship between a model and an instance of its deployment. A deployed model serves the role of the Inference Execution Instance of Figure 1 in that it captures information

about the behavior of inference server in use.

Persistent storage is through a graph database (Neo4j). Graph databases enable constant-time edge traversals and multi-hop lineage queries without expensive join patterns, which is attractive for the type of provenance relationships that are represented in our model cards. Patra’s model card search endpoint utilizes Neo4j’s full-text index to search for model card nodes by a text prompt and rank results by relevance score. The same query can be replicated in relational database management systems like PostgreSQL and MySQL.

It should be noted that it is a design decision that we make to use graph nodes to represent larger conceptual objects, and turn to properties of nodes for the attributes (details) to describe the nodes. As we will later see, queries are simpler as a result of this decision. Retrieving a model card requires a handful of node-level retrievals. But embedded properties poses an additional burden on interpretation as we will discuss in Section VI.

**FAIR Signposting.** FAIR Signposting [22] is endpoint retrieval of minimal information about objects using just a REST HEAD request which is lightweight especially for large return result sets. An AI/ML model has a unique ID that is frequently a URL pointing to the model in the repository (e.g., HuggingFace). The Patra Model Card unique ID is a unique string made up of author, model name, and version fields.

The model card ID can be used to navigate the model’s metadata using *Link-Set* endpoints. For instance, a HEAD request to `/modelcard/<ID>` returns HTTP Link headers that point to the model’s semantic metadata (the “modelcard linkset”), and a GET to `/modelcard/id/linkset` returns the full linkset JSON for that model. These responses follow the Signposting conventions: they expose URLs to the Model Card itself, the container image, datasheets, and any other related artifacts. In effect, the Patra KG acts as a signposting broker: given a model’s PID, one can resolve to all connected resources in the graph via these standard APIs.

### III. USE CASE - TRAINING TO DEPLOYMENT PIPELINE

We illustrate AI/ML model cards at the edge through a use case in the ICICLE AI Institute ecosystem. As was in Figure 2 the upper right hand corner is the starting point for the pipeline in Figure 4. In the current figure, the researcher configures an experiment to run on the edge using the MLFieldPlanner tool and GUI [17]. With an experiment configured, an instance of MLProvisioner is triggered as a Tapis job.

The provisioner reads the experiment configuration and provisions the hardware. Supported hardware includes OpenStack platforms such as Chameleon [4] and edge devices with dedicated IP addresses. The configured hardware is brought up running an instance of MLEdgeServer. MLEdgeServer supports plugins that carry out functionality (including data scoring, inference, resource monitoring). These plugins communicate through ZMQ topic streaming for intra- and inter-plugin communication (and communication external to the edge server).

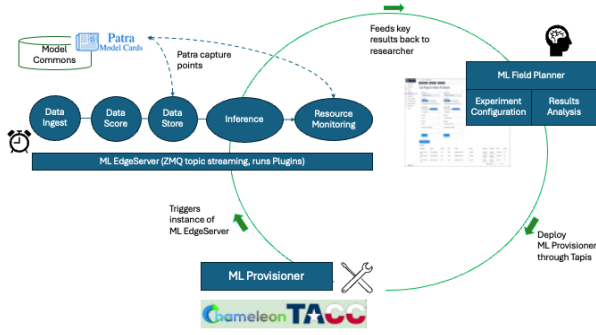


Fig. 4. Inference at the Edge through ML Workbench

The figure depicts an example pipeline for illustrative purposes. The *Data Ingest* plugin abstracts the data source (so events now use ZMQ) and may do some cleaning. The *Data Score* plugin detects shapes and label images before the event is passed to a *DataStore* plugin that chooses to discard, store, or archive images. A *Resource Monitoring* plugin captures resource (CPU, memory, energy) usage and finally, an *Inference* plugin might carry out more refined detection. The ML Field Planner listens on a topic for results that are directed at the user for results analysis.

Crucially, the data pushed to Patra includes the original model\_id provided at launch which is used to properly link edge server information to its model card.

#### IV. SYSTEM DESIGN - MODEL CONTEXT PROTOCOL

Model Context Protocol (MCP), recently released from Anthropic, offers a session-based, standardized approach to AI tool interaction that promises improved efficiency for complex, multi-step workflows. MCP standardizes the interaction between AI systems and external tools and resources through a JSON-RPC-based session protocol.

REST APIs have long served as the backbone of web services communication, providing stateless, scalable interfaces for distributed systems. Patra earlier made the highly reasonable decision to use REST for its API (Patra communication with CKN is topic-based streaming using Kafka). However, the ICICLE ecosystem could benefit from supporting MCP as a single protocol that harmonizes all data resources in the ecosystem under a single protocol. ICICLE components would thus have uniform, discoverable procedures throughout the AI/ML model lifecycle of model selection, performance tracking, provenance gathering, and lifecycle transitions, thereby streamlining automation across resource-constrained and heterogeneous settings. Unlike REST’s stateless request-response pattern, MCP maintains persistent sessions with capability discovery, enabling multi-operation workflows and contextual tool invocation. Through this session-based communication, ICICLE could enhance data resource interaction during the course of computational research workflows.

MCP declares three core primitives that servers can expose: resources, tools, and prompts. As per MCP documentation [8],

*MCP resources* are data sources that provide contextual information to AI applications (e.g., file contents, database records, API responses). *MCP tools* are executable functions that AI applications can invoke to perform actions (e.g., file operations, API calls, database queries). And *MCP Prompts* are reusable templates that help structure interactions with language models (e.g., system prompts, few-shot examples).

MCP uses the JSON-RPC 2.0 standard over lightweight, text-based transports such as HTTP(S) or Server-Sent-Events (SSE) for persistent sessions.

a) *MCP Resources, Tools in Patra*: We design Model Context Protocol (MCP) servers that serve the functionality of Patra. Our implementation extends the FastMCP framework to provide specialized functionality for the Patra Knowledge Graph, incorporating performance monitoring, caching mechanisms, and batch processing capabilities designed specifically for scientific model management scenarios.

We base our performance study on two Patra operations:

- **Model Card Retrieval**: Retrieves the full model card for a given model identifier.
- **Edge Creation**: Creates an edge (directed relationship) between two nodes in the graph. The operation validates node existence, infers relationship type from node labels via schema constraints, checks for duplicate edges, and commits the edge creation within an atomic transaction.

*Model card retrieval* is implemented through the MCP “resource” primitive because it provides contextual information about a model card in a read-only manner. Specifically, it implements the URI pattern `modelcard://{mc_id}`, where `{mc_id}` is the model card external identifier. Retrieval employs a multi-query aggregation strategy within a single Neo4j session: a Base query that retrieves the primary ModelCard node, and four independent queries fetch related entities: AI Model, Bias Analysis, Explainability Analysis, and Deployments. The results are aggregated into a nested dictionary and serialized to JSON.

We implement the *edge creation* operation that adds an edge to the graph representing a model card. This operation is implemented through the MCP “tool” primitive. `create_edge` creates directed relationships between two nodes, parameterized by Neo4j identifiers. This tool is used in the Patra-RGCN [13], an extension of Relational Graph Convolutional Networks for link prediction. Relationship types are automatically inferred from node label pairs using a static mapping.

The `create_edge` MCP tool executes a four-stage validation and commit pipeline to ensure schema compliance and idempotency: 1) The tool accepts two Neo4j `elementId` parameters and issues a Cypher query to verify both nodes exist and retrieve their label sets. 2) The tool normalizes node labels and validates the relationship against a dictionary, which encodes the knowledge graph schema as a directed adjacency structure. 3) Before creating the edge, the tool checks whether a relationship of the inferred type already exists between the two nodes. 4) If all validation passes and no duplicate



exists, the tool commits the relationship within a single Neo4j transaction.

## V. PERFORMANCE EVALUATION

To better understand the benefits of MCP for our uses, we develop two MCP servers for comparison against the existing Patra REST API. We execute end-to-end requests using two popular Patra calls: model card retrieval (get a model card) and model card search. The three server versions are thus:

- *REST*: Client invokes the Patra REST API directly via HTTP, with each request establishing a stateless connection.
- *Native MCP*: the MCP server executes queries directly against the Neo4j database using the Bolt protocol, bypassing the REST layer entirely.
- *Layered MCP*: MCP server interacts through the Patra REST layer, translating calls from the MCP tools to their corresponding REST endpoints.

Where MCP is involved, both functions are implemented as MCP tools in Python, following the standard `@mcp.tool` decorator pattern. The *Native MCP* implements the same logic as the REST endpoint does with respect to requiring multiple database queries but then wraps the response in JSON-RPC 2.0 message framing. *Layered MCP* adds a third layer: the MCP tool internally calls the REST endpoint which introduces a HTTP client-server handshake and duplicate JSON serialization (once for the REST response, again for the MCP JSON-RPC envelope).

*a) Test Environment:* We carry out two experiments: a microbenchmark experiment and a real world experiment. The microbenchmark experiment uses two nodes from the Jetstream2 system [2]. The client node is a m3.quad instance with 4 CPU, 15 GB RAM, and 20 GB hard drive. The server node is a m3.medium instance with 8 CPU, 30 GB RAM, and 60 GB hard drive. All instances run Ubuntu 24.04.3 LTS operating system and are hosted in Indiana USA. The Jetstream2 compute nodes are connected through a 100 Gbps network. The multiple runs of our experiment show no perceptible network interference.

For the real world experiment, the client node is of the same size but resides in Hawaii USA. The server node is of the same size and location in Indiana USA. All instances run Ubuntu 24.04.3 LTS operating system.

The Model Context Protocol (MCP) server and client are implemented using the MCP[CLI] library (v1.10.1). All experiment scripts utilize the MCP Python SDK for establishing Server-Sent Events (SSE) transport sessions. Each benchmark uses asynchronous I/O routines specifically Python’s `asyncio` module to minimize client-side variability. The graph database is Neo4j v5.21.0.

The server components run as isolated Docker containers (Docker Engine v28.2.2) with explicit CPU and memory limits to reduce noisy-neighbor effects and improve reproducibility. Neo4j is allocated 6 vCPU and 12 GB memory; all API servers run with 4 vCPU and 4 GB memory. The containers communicate over a Docker bridge network.

*b) Graph database:* For the microbenchmark experiment, we use model cards of a few KB in size. The graph database for the real-world experiment consists of 10,000 Deployment nodes, 1000 Experiment nodes, and 100 Device nodes. This is pseudo-synthetic data. Based on real model cards, we extended the deployment, experiment, and device information. For the wide area experiment, a single model card including its deployment information averages 13.63 MB (measured in its JSON format.)

### A. Experiment I: Microbenchmarks

The first experiment we undertake is a microbenchmark to capture both end-to-end and component latency. The layered MCP interface wraps existing REST endpoints through JSON-RPC translation, the native MCP implementation invokes data access logic directly through the FastMCP framework. Each benchmarked operation executes over a persistent SSE session and is repeated 1,000 times for statistical consistency. The experiment API endpoints return small payloads, on the order of KBs.

The total query time for retrieving (few KB) model card is  $5.8 \pm 0.7$  ms for model card retrieval operation. Model card retrieval is actually several queries, as shown in Table I. This represents the theoretical lower bound for query execution time

TABLE I  
DATABASE RETRIEVAL TIME

Node	Latency (ms)
Model Card	1.19
Model	1.0
Bias Analysis	1.04
XAI Analysis	0.95
Deployments	1.68
Total	5.87

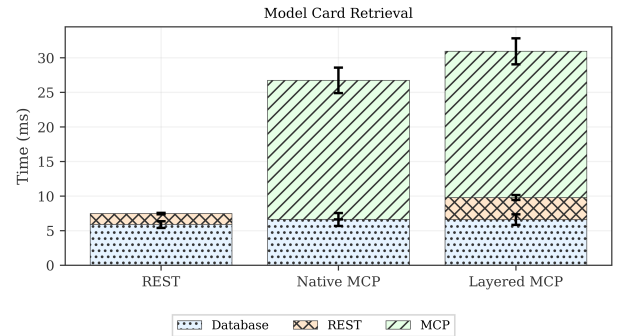


Fig. 5. Model card retrieval, few KB model card sizes

In comparing across the three server variants for the model card retrieval operation (for KB model cards), as shown in Figure 5, REST has the most efficient protocol implementation, with 7.5 ms total retrieval time. Its small deviation reflects Flask’s mature HTTP handling and direct JSON serialization paths.

The Native MCP server has a 3.6 times higher turnaround time than the REST server does for model card retrieval with

26.7 ms. This performance differential stems from MCP’s JSON-RPC 2.0 protocol complexity, including message envelope construction, tool dispatch mechanics, and SSE session management [3]. The layered REST+MCP architecture is 4.1 times the turnaround as does vanilla REST; this exposes a double serialization penalty: the initial Flask JSON encoding followed by MCP JSON-RPC envelope wrapping. The additional REST layer for the REST+MCP is a 16% overhead from native MCP. While operationally convenient for maintaining backward compatibility, the layered architecture’s 4× overhead relative to REST establishes it as suitable primarily for development environments and transitional deployments rather than production systems optimizing for performance.

The difference between native and layered MCP architectures centers on protocol integration depth and data flow abstraction. In a native implementation, the Model Context Protocol (MCP) interacts directly with the application layer, bypassing any intermediary HTTP-based abstraction such as REST. This design aligns with direct binding theory where minimal protocol layering leads to reduced serialization overhead and tighter coupling between semantic logic and transport mechanisms [3]. Native MCP servers perform all operations within the MCP tool abstraction, handling JSON-RPC message parsing, state management, and response serialization directly. The resulting simplicity in data flow can reduce latency but requires re-implementation of existing REST logic and error-handling layers, increasing development complexity. In contrast, the layered MCP architecture applies an adapter pattern where MCP acts as a proxy wrapping existing REST endpoints. This approach leverages REST’s mature API contracts and established tooling while exposing equivalent functionality through MCP’s tool and resource schemas.

Layering preserves backward compatibility and accelerates adoption but introduces double serialization costs and compounded transport overhead. The choice between the two approaches, therefore, reflects a tradeoff between architectural purity and integration pragmatism—native designs favor performance and stability within agentic infrastructures, while layered designs favor interoperability and evolutionary system transitions.

In wide-area scenarios, connection time escalates, suggesting that SSE connection setup involves multiple network round-trips. The server-side Neo4j execution time is identical to the REST implementation. In Native MCP, database latency is comparable to that of REST, validating that protocol overhead does not impact backend query performance. Server Processing latency includes MCP session management, response framing (SSE encoding), JSON serialization, and accumulated delays from the persistent connection model. In a Layered MCP implementation, the REST overhead is attributed to the additional cost of marshaling MCP operations into HTTP requests and sending them to the REST API. For edge creation, REST overhead is constant, representing HTTP serialization and deserialization, as well as network transmission to the REST layer. For model card retrieval, REST overhead becomes substantial, indicating that transmitting large payloads through

the REST layer incurs measurable latency beyond direct database access.

## B. Experiment II: Real World

In this second experiment we drive towards a more realistic scenario and do so in a few dimensions. As indicated earlier, Patra is unique in that it accumulates usage information about model inference servers. As such, model cards can grow to be sizeable. In this experiment we use synthetic model cards that are 13 MB in size (when formatted as a JSON object).

We further create a second client that is located on Jetstream2 servers located in Hawaii (with the Patra server still residing in Indiana). We evaluate the two operations using substantially larger model cards and introducing a wide area dimension to the testing.

With large model cards (13.63 MB), the database query response time grows to 7843 ms, see Table II. This compared to the retrieval time of a small (few KB) model card which is under 6 ms.

TABLE II  
MODEL CARD RETRIEVAL; LARGE MODEL CARD

Node	Latency (ms)
Model Card	7843.83

In measuring client-server end-to-end latency, timing is broken down into the following components:

- *Connection Setup*: captures the time required to establish a connection between the client and the server. It includes the TCP handshake and the creation of a persistent HTTP connection for event streaming. Connection setup is a one-time cost and is analogous to the initial connection establishment in traditional HTTP protocols.
- *SSE Handshake* Server-Sent Events (SSE) is a transport session connection between the MCP client and server. It involves socket initialization and HTTP upgrade negotiation. Local connection overhead includes the initialization of the MCP session. MCP experiment scripts utilize the MCP Python SDK for establishing Server-Sent Events (SSE) transport sessions. SSE is used in MCP but not REST to maintain the session between requests (REST is stateless).
- *Server Processing*: the time taken to issue a resource read command over the established session and receive the response. It includes network transfer, protocol overhead, server-side processing, and backend database access culminating in the delivery of the requested resource back to the client.

The REST results (shown in Figure 6, show connection setup times that are fairly independent of the distance between client and server. As server processing includes data transfer back to the client, the wide area numbers reflect the large payload being returned. Note the log scale used in all of Figures 6, 7 and 8.

In comparing native MCP to layered MCP for Model Card Retrieval and the large model card (Figures 7 and 8

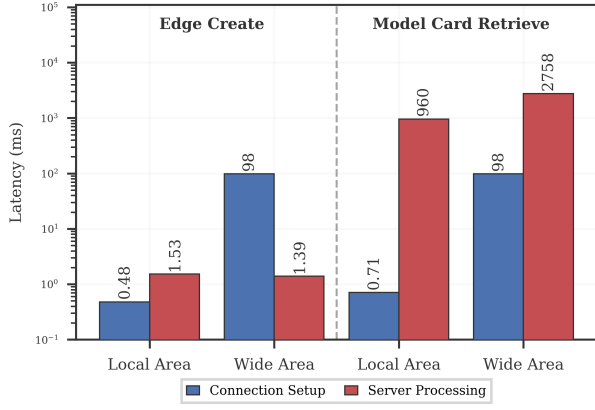


Fig. 6. REST, large model card

respectively), one can observe that with REST set aside in Figure 8 the latencies track one another pretty closely. What this suggests is that the use of REST does not particularly penalize MCP. It does, however introduce its own overheads that are evident in purple in Figure 8. In comparing REST (Figure 6 and native MCP 7, one can see sizeable overheads in the wide area setting for Edge Create. Edge Create returns few KB result sets, so there is additional protocol overheads with MCP in all of connection setup, SSE Handshake, and Server processing. MCP is 7 times slower than REST for this particular type of operation.

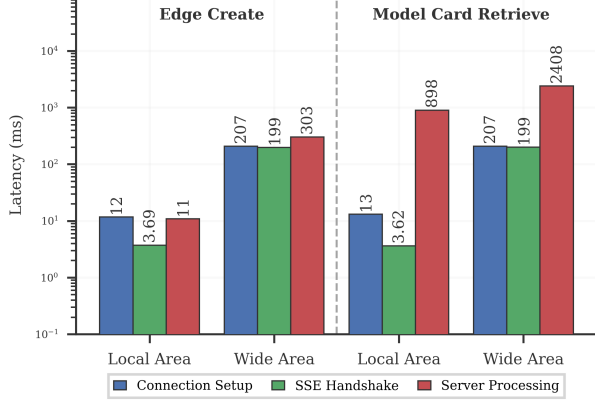


Fig. 7. Native MCP, large model card

We conclude in Figure 9 by plotting total end-to-end latencies for the local versus wide area configurations and the three server versions (REST, native MCP, and Layered MCP). Note the log scale on the Y axis. What this shows is that when result sets are large, the difference between Native MCP and Layered MCP, and even including REST, are not that large. This holds even more true when wide areas are involved.

## VI. DISCUSSION

One of the most interesting outcomes of the quantitative evaluation is the additive nature of REST + MCP. That is, when one layers REST + MCP, one pretty well takes on the

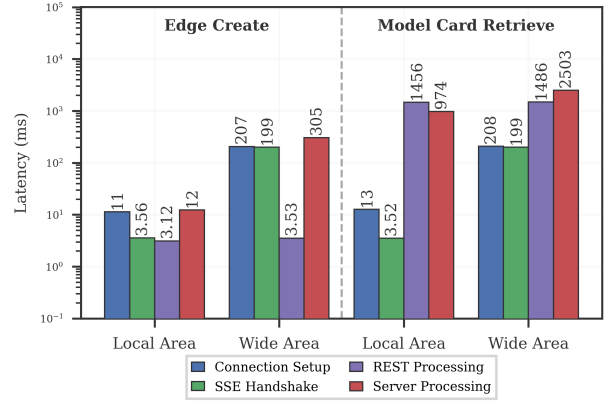


Fig. 8. Layered MCP, large model card

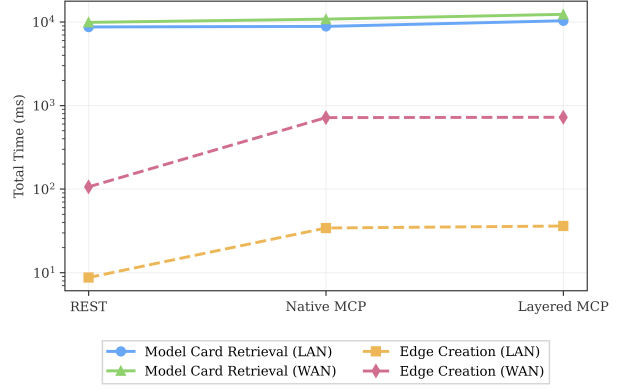


Fig. 9. Total turnaround time for each approach on a log scale.

overheads of both protocols. But in the broader setting of big result sets, non-trivial database graph walks, and longer distances, the significance of the additional overhead diminishes in importance.

A fundamental decision point with MCP adoption boils down to whether a data-holding service can take advantage of MCP's session orientation which is supported through the Server Side Events (SSE) protocol. SSE exists to support server-side notifications, which give a server the mechanism to push out notifications/events over time to a listening client, often an AI Agent.

Notifications imply the existence of a client that wants to know about change over time in something. Suppose, for example, an AI Agent maintains a internal model of the usage of a particular AI/ML model or set of models over time. Using the historical context that an LLM could process, the agent could flag usage that deviates from historical use. The role of the Patra repository then is to notify the agent each time there is a new deployment of a model card.

More broadly this interaction between data store and AI agent is cast as an interaction method such as the ReAct (Reson + Act) method [7]. In this method, an agent presents the state of the world to an LLM, solicits LLM response, then asks the LLM what it wants to do next. An agent begins



by writing a thought about the given task. The agent then performs an action based on that thought, and the output is observed. This cycle can be repeated until the task is complete.

As Masterman *et al.* [7] point out, however, the ReAct method is not without its limitations. "While intertwining reasoning, observation, and action improves trustworthiness", they state, "the model can repetitively generate the same thoughts and actions and fail to create new thoughts to provoke finishing the task and exiting the ReAct loop."

Circumventing the repetitive loop of the same thoughts and actions from the LLM agent could be accomplished by incorporating human feedback during the execution of the task.

We turn back to the initial argument in the paper: what does it mean to benchmark model use at points during its lifecycle rather than just at training time. We interpret benchmarking model use to having a "by whom" component and a "how" component.

To the former, "by whom", we are cognizant that model cards belong to the person who created the model. There are many implications of model card ownership that we do not go into here, but before describing how we track the creator, we point out that an individual's organizational affiliation is a strong indication of how a model has been used. If a model is used by a county courthouse, this tells us something about its use. Organizational affiliation is a placeholder for us - an indicator of use that we intend to push on more deeply.

Model Cards have the same creator as the AI/ML model by the following: using the toolkit's `submit()` method, the user can simultaneously upload the model artifacts to a selected repository (e.g., HuggingFace or GitHub) and ingest the Model Card to the Patra server. Ownership is established through the Tapis hosted infrastructure API. Tapis provides the authorization and authentication framework for the ICICLE cyberinfrastructure, maintaining the identity of individual researchers using the platform. A researcher can, for instance, retrieve a trained model artifact using their credentials or share it with a collaborator via a time-limited access token. The integration ensures that an actual user of the ICICLE environment is associated with each model card in the system.

As to the "how", as captured in the data model of Figure 3, our group is exploring the capture and representation of information surrounding the use of a model in the form of the experiment (or task graph in agentic AI). For the model card to have the needed information to deploy the right model we think will give us leverage, along with the "by whom" to trigger alerts when a model is being used outside of some tolerance of its deployment in the past. This is future work for us.

## VII. RELATED WORK

The Patra Knowledge Base runs as a persistent service, ingesting JSON model cards as graph representations using, in part, the PROV-ML ontology [1]. Several recent projects in AI MLOps and distributed systems have begun to adopt agent-based architectures, but only a few leverage open protocols for tool/resource discovery. MCP advances the state of the art

by formalizing the way agents communicate, interrogate, and control external resources (such as artifact registries, policy engines, and provenance graphs) in a vendor- and language-neutral manner. Within Patra and CKN, this means that model lifecycle, deployment, and provenance steps may be accessed and controlled by any agent that speaks MCP, simplifying integration with major AI ecosystems or federated scientific infrastructures.

Protocol performance evaluation has been extensively studied in the context of web services. Pautasso *et al.* [11] established foundational principles for REST API design and performance characteristics. Recent work by Zhang *et al.* [26] compared gRPC, REST, and GraphQL across e-commerce workloads, finding significant performance variations based on operation complexity and data access patterns. However, existing studies focus primarily on traditional web service scenarios. The unique characteristics of AI model management—including complex metadata structures, multi-criteria decision making, and edge deployment constraints—have not been systematically evaluated across different protocol architectures.

Comparable efforts in reproducibility and provenance-aware scientific workflows include the Sciunit framework by Tanu Malik *et al.* [19], which encapsulates scientific processes as reusable research objects with captured provenance for transparent re-execution. Recent works leverage large language models to automate documentation using pipelines like Card-Gen [6].

LinkEdge [12] is a lightweight platform that bridges cloud MLOps pipelines with edge devices. The open, modular nature of LinkEdge aligns with Patra's goals of scalable and transparent edge-cloud ML management. In our architecture, Patra issues globally unique PIDs for every serialized model and version of its ModelCard, while TAPIS [18] is the underlying object store and resolution fabric. Patra Model Cards employ signposting. The seminal work of Van de Sompel *et al.* [22] introduces FAIR Signposting Profile as implementation guidelines for exposing machine-actionable navigation links using standardized HTTP headers and HTML link elements. Soiland-Reyes *et al.* [16] demonstrates the combination of signposting with RO-Crate as a practical, web-centric approach to implement FAIR Digital Objects. Kreuzberger *et al.* [5] point out that MLOps platforms often rely on internal identifiers. Internal identifiers are inconsistent with the principles of FAIR software and data and open science.

Renan *et al.* [15] articulates the vision for using workflow provenance across the edge-cloud continuum to support Responsible and Trustworthy AI. They argue that provenance data is critical but note a gap in current systems where provenance management is not well-integrated with heterogeneous infrastructures and ML life-cycles. Venkataramanan *et al.* [21] demonstrate the scalability of using Knowledge Graphs to integrate diverse ML metadata from multiple sources for querying and recommendation.

## VIII. CONCLUSION AND FUTURE WORK

In this article we contextualize model cards within the larger setting of the ICICLE AI Institute software ecosystem. We quantify overheads of REST versus Model Context Protocol (MCP) protocols in serving model cards in various settings. We call out the most significant distinction between REST and MCP, and that is the session orientation of the latter. This we explore in terms of agentic AI and whether Patra model cards' goal of AI accountability can be reached by its support of benchmarking model use at points in model lifecycle (beyond its initial training). We are encouraged.

An additional open question addresses Patra's heavy dependency on attributes/constraints attached to nodes and tasking of REST endpoints with sole knowledge of schema information (such as what defines a model card). These are two major design choices that need further study in light of AI interactions with the repository.

## IX. ACKNOWLEDGEMENTS

We thank the entire ICICLE Smart Fields team, but particularly Christopher Stewart of Ohio State University.

## REFERENCES

- [1] Khalid Belhajjame, Reza B'Far, James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, Timothy Lebo, Jim McCusker, et al. PROV-DM: The prov data model. *W3C Recommendation*, 14:15–16, 2013.
- [2] David Y. Hancock, Jeremy Fischer, John Michael Lowe, Winona Snapp-Childs, Marlon Pierce, Suresh Marru, J. Eric Coulter, Matthew Vaughn, Brian Beck, Nirav Merchant, Edwin Skidmore, and Gwen Jacobs. Jetstream2: Accelerating cloud computing via jetstream. In *Practice and Experience in Advanced Research Computing*, PEARC '21, New York, NY, USA, 2021. Association for Computing Machinery.
- [3] Tyler Jackson, Liying Zhang, and David Berman. Streaming technologies and serialization protocols: A comparative performance study for distributed systems. *arXiv preprint arXiv:2407.13494*, 2024. Available online: arXiv, July 2024.
- [4] Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Collieran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. Lessons learned from the Chameleon testbed. In *Proc 2020 USENIX Annual Technical Conf (USENIX ATC '20)*. USENIX Association, July 2020.
- [5] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. Machine learning operations (mlops): Overview, definition, and architecture. *IEEE Access*, 11:31866–31879, 2023.
- [6] Jiarui Liu, Wenkai Li, Zhijiang Jin, and Mona Diab. Automatic generation of model and data cards: A step towards responsible AI. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proc 2024 Conf North American Chapter of the Assn for Computational Linguistics: Human Language Technologies (Vol 1)*, pages 1975–1997, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
- [7] Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey, 2024.
- [8] MCP. Model context protocol. <https://modelcontextprotocol.io/docs/getting-started/intro>, 2024.
- [9] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proc of Conf on Fairness, Accountability, and Transparency*, pages 220–229, 2019.
- [10] Dhableswar K. Panda, Vipin Chaudhary, Eric Fosler-Lussier, Raghu Machiraju, Amit Majumdar, Beth Plale, Rajiv Ramnath, Ponnuswamy Sadayappan, Neelima Savardekar, and Karen Tomko. Creating intelligent cyberinfrastructure for democratizing AI. *AI Magazine*, 45(1):22–28, 2024.
- [11] Cesare Pautasso and Erik Wilde. Restful web services: principles, patterns, emerging technologies. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, page 1359–1360, New York, NY, USA, 2010. Association for Computing Machinery.
- [12] Ella Peltonen and Savidu Dias. Linkededge: Open-sourced mlops integration with iot edge. In *Proceedings of the 3rd Eclipse Security, AI, Architecture and Modelling Conference on Cloud to Edge Continuum*, eSAAM '23, page 67–76, New York, NY, USA, 2023. Association for Computing Machinery.
- [13] Krishna Priya, Sachith Withana, and Beth Plale. Patra-rgcn: Missing link prediction in model card graphs through node property encodings. In *2025 IEEE International Conference on eScience (eScience)*, pages 131–140, 2025.
- [14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.
- [15] Renan Santos Souza, Silvina Caino-Lores, Mark Coletti, Tyler Skluzacek, Alexandru Costan, Fred Suter, Marta Mattoso, and Rafael Ferreira Da Silva. Workflow provenance in the computing continuum for responsible, trustworthy, and energy-efficient ai, 09 2024.
- [16] Stian Soiland-Reyes, Peter Sefton, Simone Leo, Leyla Jael Castro, Claus Weiland, and Herbert Van de Sompel. Practical webby fdos with ro-crate and fair signposting: Experiences and lessons learned. *Open Conference Proceedings*, 5, March 2025.
- [17] J. Stubbs, S. Balasubramaniam, and et al. ML Field Planner: Analyzing and optimizing ML pipelines for field research. In *Proc. Practice and Experience in Advanced Research Computing 2025 (PEARC 2025)*. ACM, 2025.
- [18] Joe Stubbs, Richard Cardone, Mike Packard, Anagha Jamthe, Smruti Padhy, Steve Terry, Julia Looney, Joseph Meiring, Steve Black, Maytal Dahan, Sean Cleveland, and Gwen Jacobs. Tapis: An api platform for reproducible, distributed computational research. In Kohei Arai, editor, *Advances in Information and Communication*, pages 878–900, Cham, 2021. Springer Int'l Publishing.
- [19] Dai Hai Ton That, Gabriel Fils, Zhihao Yuan, and Tanu Malik. Sciunits: Reusable research objects. In *IEEE 13th Int'l Conf on e-Science (e-Science)*, pages 374–383, 2017.
- [20] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [21] Revathy Venkataramanan, Aalap Tripathy, Martin Foltin, Hong Yung Yip, Annmary Justine, and Amit Sheth. Knowledge graph empowered machine learning pipelines for improved efficiency, reusability, and explainability. *IEEE Internet Computing*, 27(1):81–88, 2023.
- [22] Mark Wilkinson, Susanna-Assunta Sansone, Marjan Grootveld, Richard Dennis, David Hecker, Robert Huber, Stian Soiland-Reyes, Herbert {Van de Sompel}, Andreas Czerniak, Milo Thurston, Allyson Lister, and Alban Gaignard. *Report on FAIR Signposting and its Uptake by the Community: EOSC Task Force on FAIR Metrics and Data Quality*. Zenodo, Switzerland, January 2024.
- [23] Sachith Withana and Beth Plale. CKN: An Edge AI Distributed Framework. In *2023 IEEE 19th Int'l Conf on e-Science (e-Science)*, pages 1–10. IEEE, 2023.
- [24] Sachith Withana and Beth Plale. Patra modelcards: AI/ML accountability in the edge-cloud continuum. In *2024 IEEE 20th Int'l Conf on e-Science (e-Science)*, pages 1–10. IEEE, 2024.
- [25] M. Brundage et al. Y. Shavit, S. Agarwal. Practice for governing agentic AI systems, 2024.
- [26] Liang Zhang, Kaiyu Pang, Jie Xu, Hongliang Chen, Yifan Wang, and Xin Liu. High performance microservice communication technology based on modified remote procedure call. *Scientific Reports*, 13(1):12141, 2023.