Automation Pipeline for ML-Assisted Scientific Data Curation and Discovery

Nishan Pantha[1], Bishwas Praveen[1], Muthukumaran Ramasubramanian[1], Sajil Awale[1], Simran KC[1], Carson Davis[1], Iksha Gurung[1], Kaylin Bugbee[2], Manil Maskey[2], Rahul Ramachandran[2]

[1] - The University of Alabama in Huntsville (UAH)
[2] - National Aeronautics and Space Administration (NASA)

National Aeronautics and Space Administration

AI-Assisted Curation for Scientific Discovery

Our team developed an **AI-assisted automation pipeline** (*Figure B*) to streamline curation at scale for the **Science Discovery Engine (SDE)** — combining ML, human expertise, and infrastructure orchestration. *(SDE provides a platform for scalable scientific content discovery)*.

This **end-to-end automated content curation pipeline** supports:

- ❑ **Human-in-the-loop evaluation of ML curated tags**
- ❑ **Generalized Inference pipeline speeding up subsequent modelling efforts**
- ❑ **Domain-adapted classification models that perform better than generic counterparts**

Human-in-the-Loop: SME-Driven Iterative Refinement

A structured HITL process ensures ML outputs reflect SME knowledge and maintain domain fidelity.

1. SME-Guided Scope Definition: SMEs define classification goals, edge cases, and quality standards. This establishes alignment between ML predictions and domain needs.

2. Expert-Labeled Data Collection: Curated training data forms the foundation for robust models.

3. Active Labeling + Iterative Feedback Loop: SMEs validate model predictions, flag misclassifications, and improve dataset quality via error analysis *(Figure A)*.

4. Continuous Model Evolution: Weak areas are identified via strategic sampling; retrained models are redeployed, closing the feedback loop.

Outcome: Adaptive ML models continuously refined through real-world SME insight

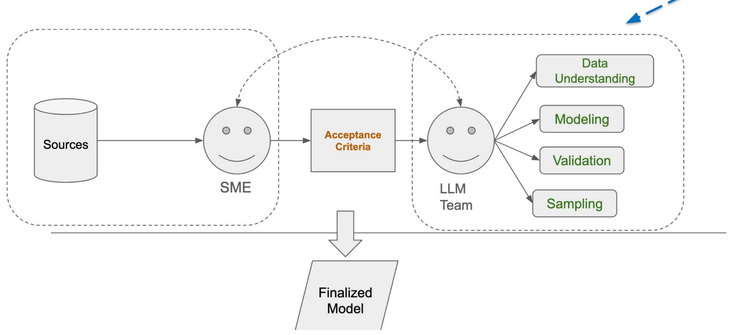

***Figure A**: Iterative model refinement process that includes continuous communication between SMEs, stakeholders, and the LLM team*

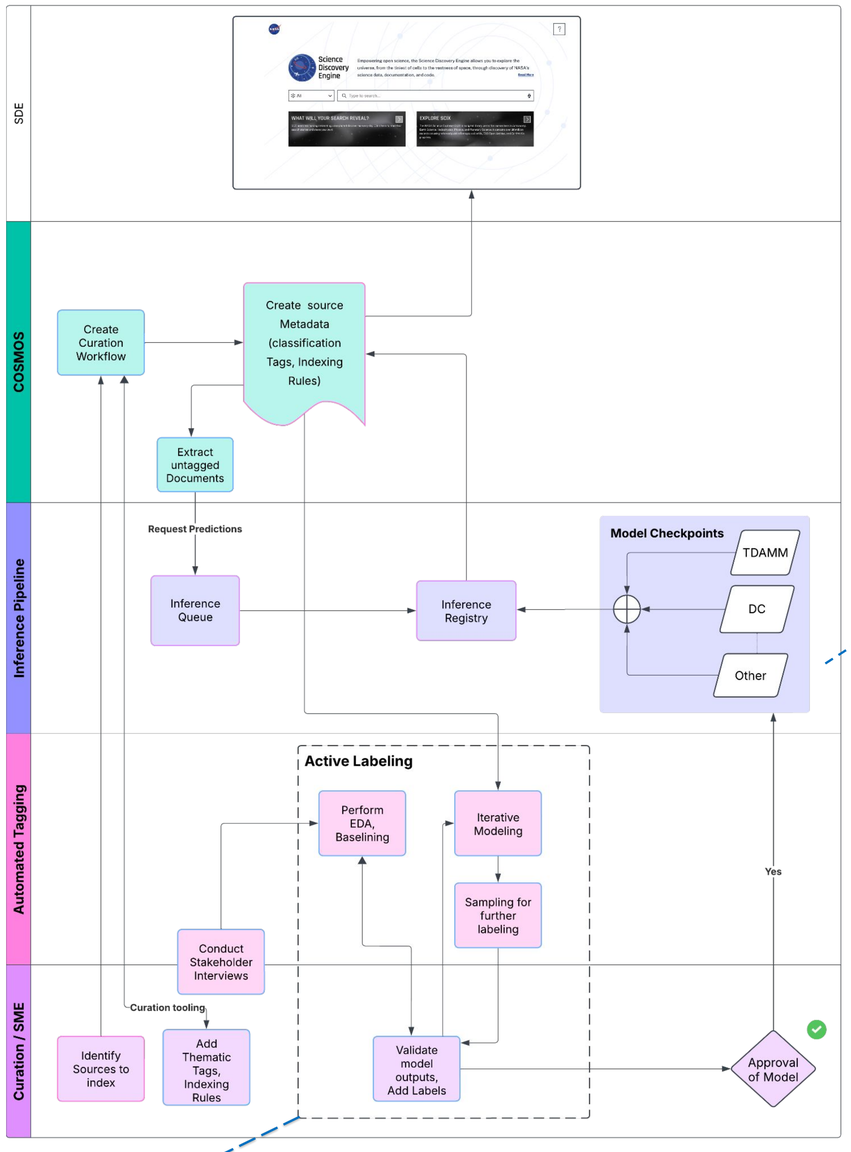

***Figure B**: SDE curation workflow with AI-assisted tagging and inference pipeline in the loop.*

ML Inference Pipeline

An ML inference pipeline is composed of a core orchestration engine integrated with ML models (Figure C), deployed into SDE's internal curation platform, COSMOS. Features include:

- ❑ Async REST API via FastAPI
- ❑ Redis-backed job queuing
- ❑ Dockerized deployment
- ❑ Extensible ML inference with plug-n-play model registry

COSMOS uses this to trigger predictions, track results and stats, thus helping in curation workflow.

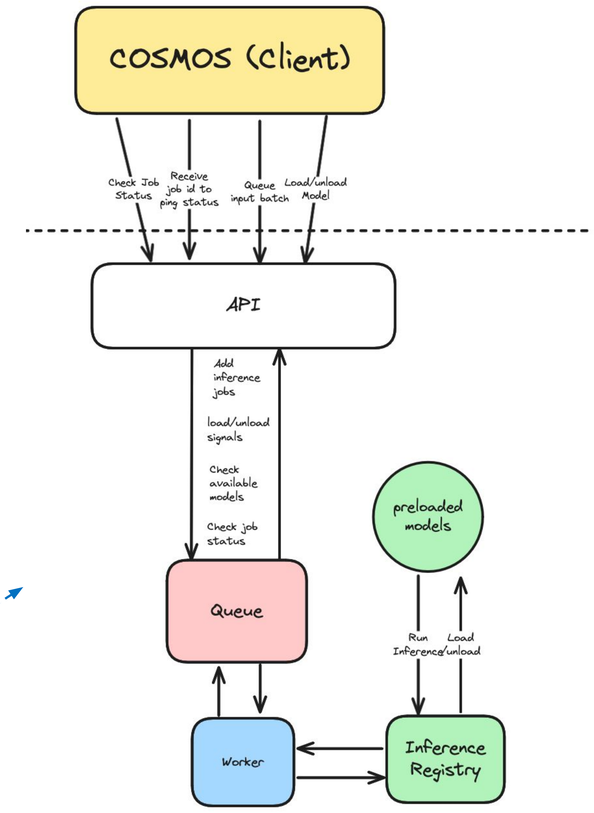

***Figure C**: A schematic diagram of the inference pipeline architecture that shows the overall flow for inference*

Production Use Cases

While the framework is domain-agnostic, two successful integrations illustrate its power (with more being internally tested):

- ❑ **Time Domain and Multi-Messenger Astrophysics (TDAMM)**: Tags astronomy documents into 36 phenomena.
- ❑ **Division Classifier (DC)**: Tags document across NASA's five science divisions: Earth Science, Astrophysics, Biological & Physical Science, Heliophysics, and Planetary Science (>96% precision across divisions).

TDAMM Classifier ***Division Classifier***

This work is supported by NASA Grant 80MSFC22M004