Harnessing Large Language Models for Scientific Endeavors

Rahul Ramachandran[1], Manil Maskey[1], Kaylin Bugbee[1], Mike Little[3], Elizabeth Fancher[6], Muthukumaran Ramasubramanian[5], Bishwaranjan Bhattaacharjee[4], Raghu Ganti[4], Avi Sil[4], Lauren Sanders[7], Sylvain Costes[7], Sergi Blanco-Cuaresma[8], Kelly Lockhart[8], Thomas Allen[8], Felix Grazes[8], Megan Ansdell[2], Alberto Accomazzi[8], Tsendgar Lee[12], Sanaz Vahidinia[2], Ryan McGranaghan[9], Armin Mehrabian[3]

NASA MSFC[1], NASA HQ[2], NASA GSFC[3], IBM Research[4], UAH[5], Barrios Technology[6], NASA Ames[7], Harvard CfA[8], JPL[9]

AGU Fall Meeting, 2023

Outline

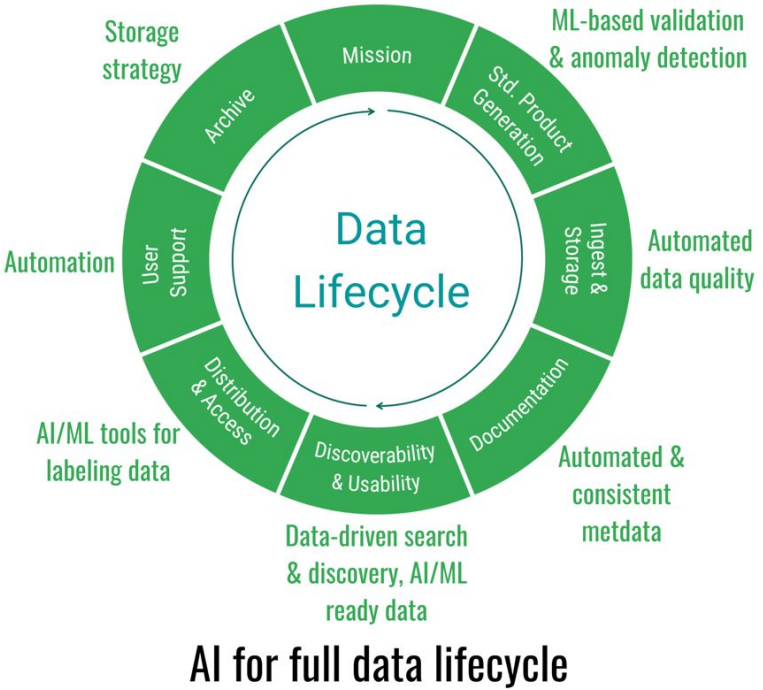
- Background
- NASA CSDO's Language Model for Science Mission Directorate
- Case study demonstrations
- Current status and Future Plans

Using LLMs to Augment Data and Research Lifecycle

LLM Infusion

- Data
  Improving existing data governance and management processes in the data lifecycle

- Analysis
  Accelerating scientific discovery and applications by simplifying use of existing data resources



Limitations of LLMs

Generalist Nature of LLMs:

- Designed to handle a wide range of topics and not specialized for any particular domain.
- Lack depth in specific scientific areas.

Limited Understanding of Scientific Concepts:

- Scientific literature often contains complex jargon, equations, and references
- LLMs might produce or propagate misconceptions or outdated information.
- Difficulty in grasping the nuanced context of certain scientific discussions.

Reasons for Limitations:

- LLMs are trained on vast and diverse datasets, but not scientific literature.
- Many scientific papers are behind paywalls, limiting their inclusion in publicly available training datasets.
- Even if scientific papers are part of the training data, the sheer volume of other non-scientific data can dilute the depth of scientific understanding.
- Science is continuously evolving, and LLMs might not be up-to-date

Language Model Initiative for SMD

- Multiple independent initiatives within NASA's SMD utilizing Natural Language Technology (NLT) for scientific research and discovery.

- Retreat:
  - Held in Washington DC on March 6, 2023.
  - Purpose: Understand the impact of Large Language Models on SMD's activities and align existing efforts.

- Event Highlights:
  - Attendance by NLT practitioners from across SMD.
  - Facilitated collaboration among attendees.
  - Focus on the design, construction, and evaluation of a common Large Language Model tailored for SMD's needs

Language Model Initiative for SMD

Key Outcomes from the Retreat:

- Domain-Specific LLM Necessity:
  - Participants concluded the need for an SMD-specific LLM.
  - Such a model would offer more technical depth than general-purpose models
  - Aimed to better support downstream science applications.

- LLMs as Augmentation Tools:
  - LLMs are meant to augment, not fully automate, processes.
  - Human oversight is essential to verify results.
  - LLMs can significantly enhance process efficiency.

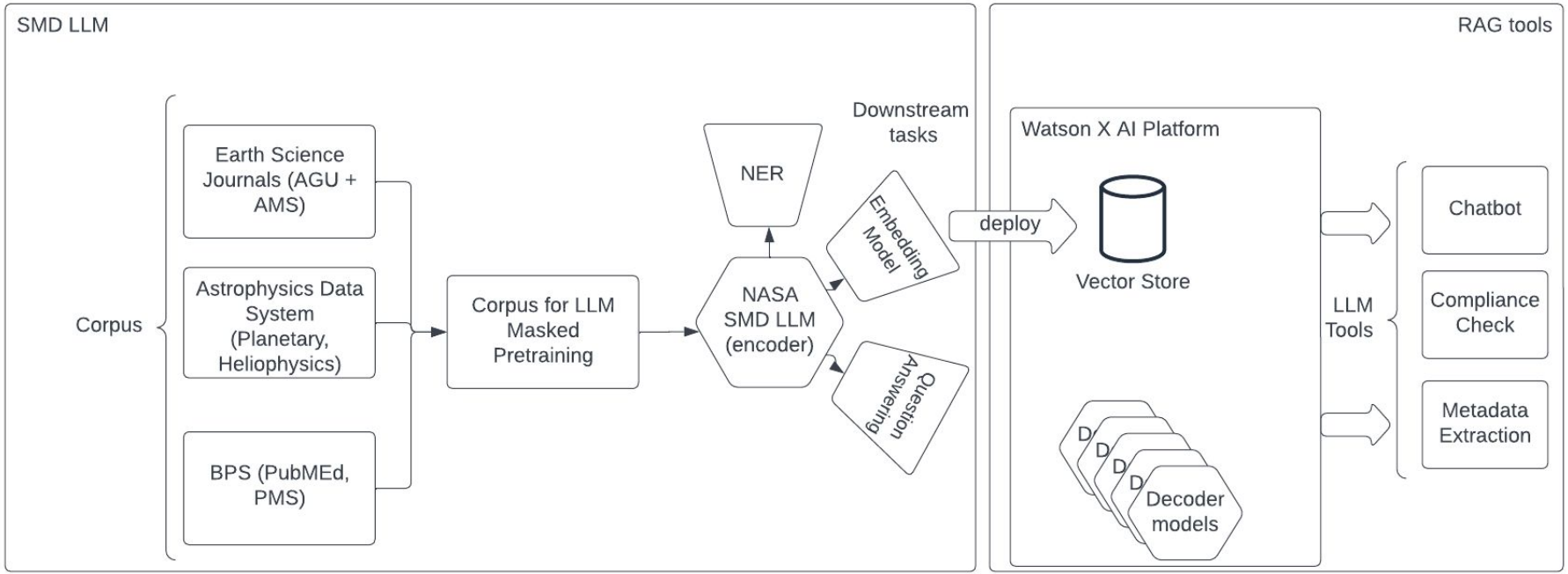
- Recommendations for CSDO:
  - Establish a focused Working Group.
  - Leverage the public-private partnership with with IBM Research for SMD LLM development.
  - Create a Validation test suite for LLM training.
  - Develop a general Evaluation Framework.
  -

Both the Validation suite and Evaluation Framework will also benefit the broader science community.

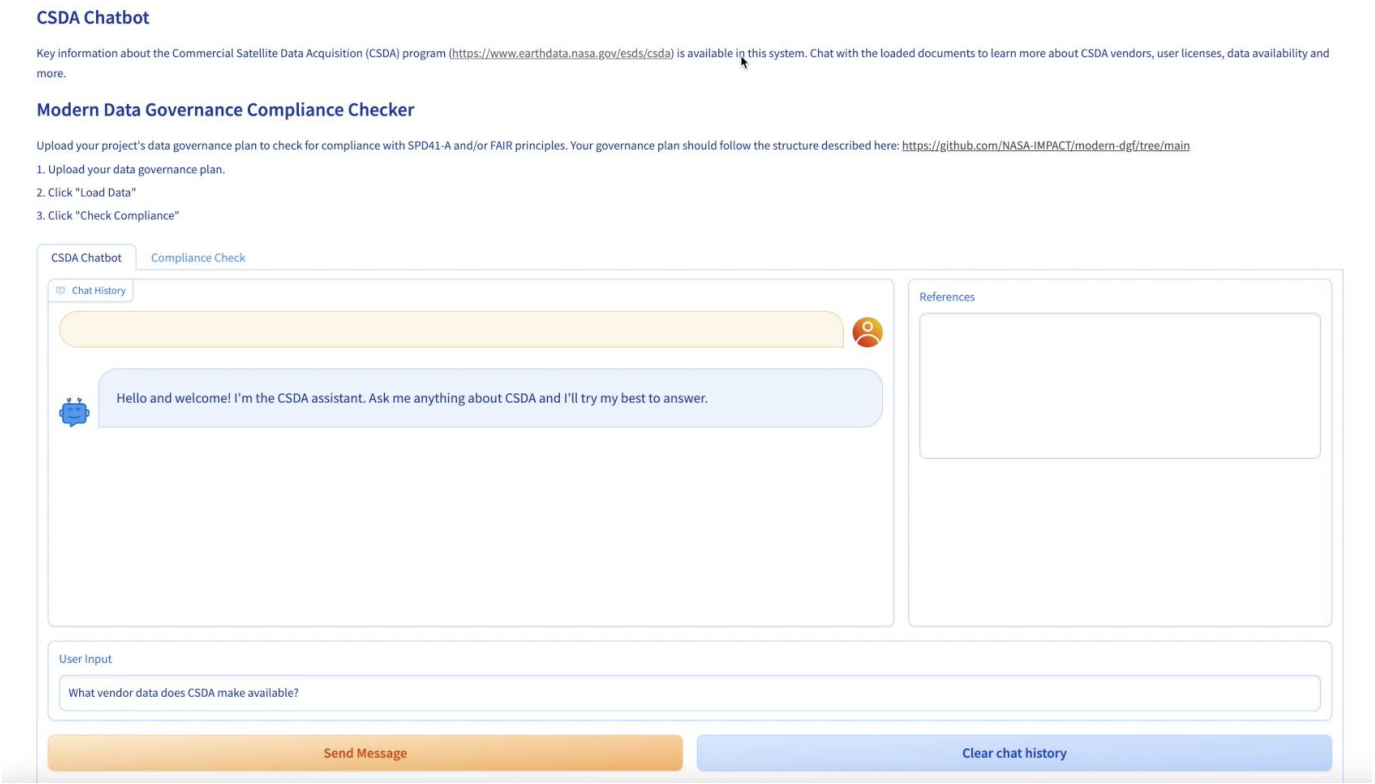
Current status

- Curated Training Corpus
  - ESD: ~120,000 full papers from AGU, AMS.
  - ADS/SciX full papers
  - Pub Med, Wikipedia
- Development of Encoder Model and Huggingface release.
- Evaluation of LLMs
- Develop Retrieval-Augmented Generation(RAG) framework that combines the capabilities of generative models with encoder based retrieval and reranking models, among other sources.
- Enable Application Development on the RAG Framework
  - Utilizing LLMs and curated content to build specialized applications.
  - Flexibility for groups to select from different LLMs for app development.
  - Deployment of IBM's Watson-x Framework on NASA secure cloud environment:
    - LLaMA 2.0 models hosted on IBM Watson-x for increased availability and throughput.
    - Embedding models based on NASA SMD LLM for increased relevance of retrieval systems

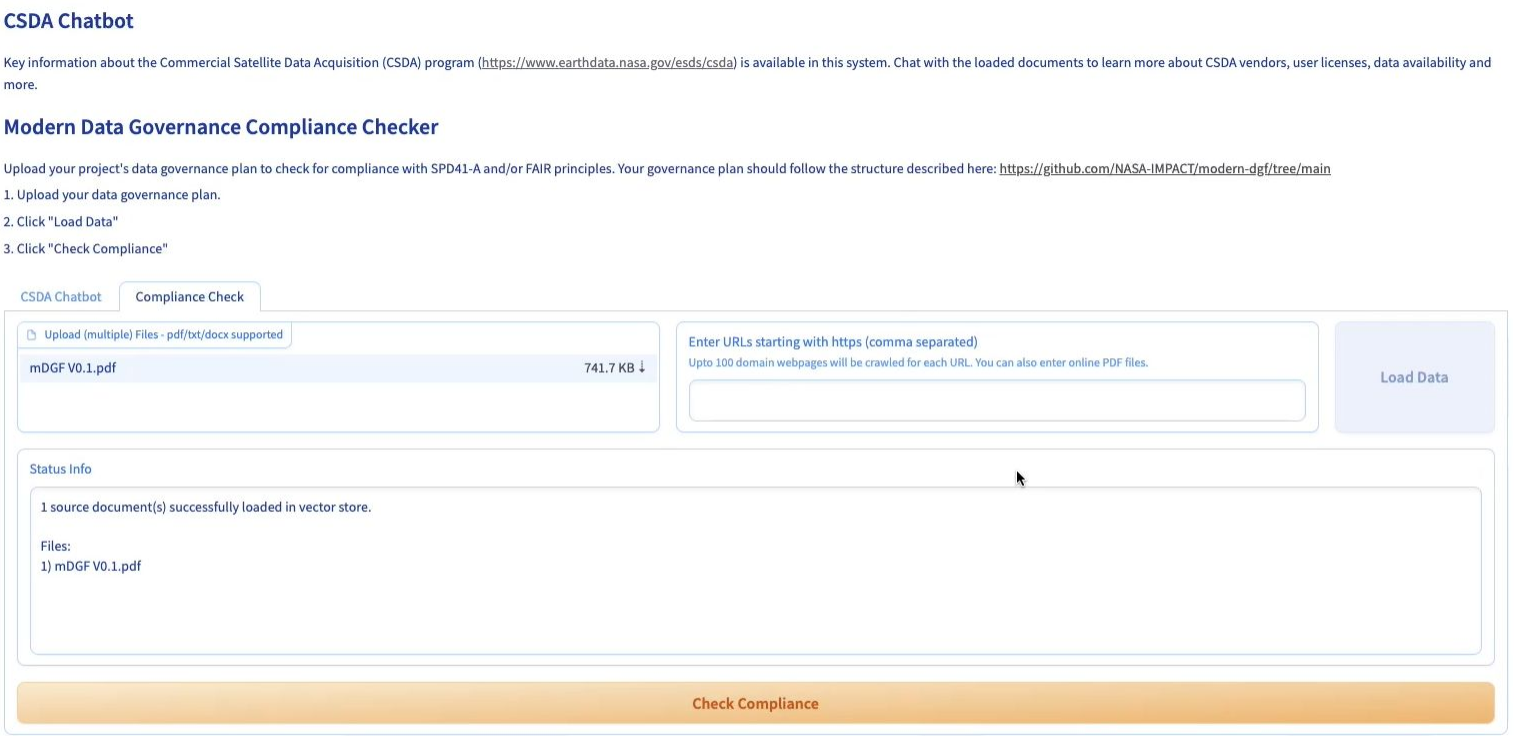
Current approach



Demo 1 - Q&A


CSDA Chatbot
Key information about the Commercial Satellite Data Acquisition (CSDA) program (https://www.earthdata.nasa.gov/esds/csda) is available in this system. Chat with the loaded documents to learn more about CSDA vendors, user licenses, data availability and more.
Modern Data Governance Compliance Checker
Upload your project's data governance plan to check for compliance with SPD41-A and/or FAIR principles. Your governance plan should follow the structure described here: https://github.com/NASA-IMPACT/modern-dgf/tree/main
1. Upload your data governance plan.
2. Click "Load Data"
3. Click "Check Compliance"
CSDA Chatbot
Compliance Check
Chat History
Hello and welcome! I'm the CSDA assistant. Ask me anything about CSDA and I'll try my best to answer.
References
User Input
What vendor data does CSDA make available?
Send Message
Clear chat history

Demo 2 - Compliance checker



Future work

- Potential for generative models for internal use
- Share of evaluation suite and performance benchmarks to the science community
- Develop educational resources on:
  - How to effectively leverage the RAG Framework.
  - Best practices for selecting and adapting different models for applications
  - Cost effective strategies to deploy models into production environments.
- More LLM applications enabled by RAG, and Metadata Extraction.
- Use case-Specific Parameter Efficient Fine Tuning (PEFT) of LLMs.

Thank you.

rahul.ramachandran@nasa.gov