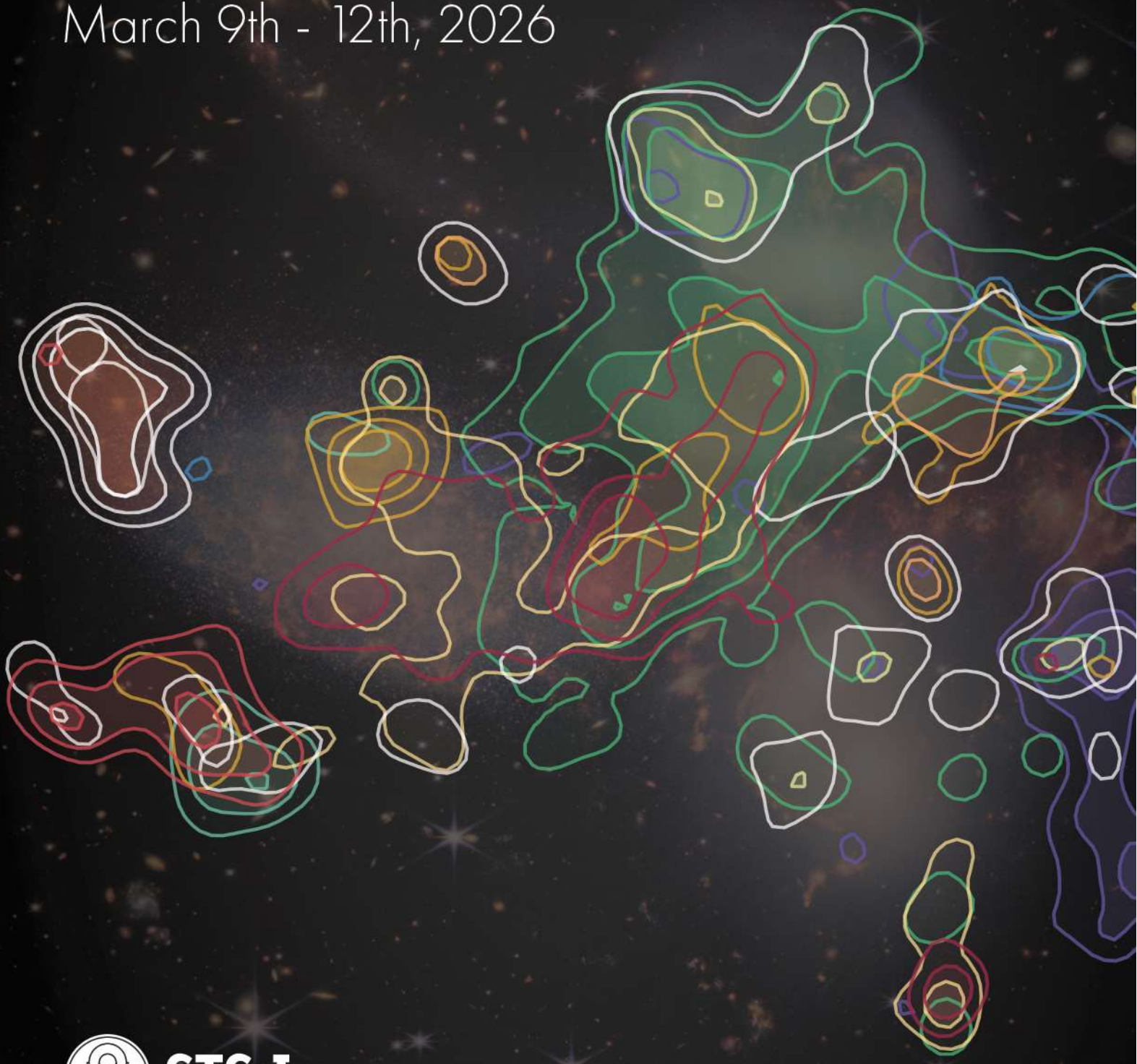
Language AI in the Space Sciences Workshop

March 9th - 12th, 2026



STScI

Table of Contents

Workshop Schedule

STScI

Language AI in the Space Sciences

March 9 – 12, 2026 at Space Telescope Science Institute – Schedule of Events

Monday, March 9, 2026

Time	Title	Speaker	Location
1:00pm	Registration		Lobby
1:35pm	Coordination for Tutorials		Auditorium
1:50pm	Tutorial Session A • "Building an AI Policy" with Michelle Ntampaka (Café Con) • "AI Agents" with Lorenz Ehrlich., Miguel Doctor Yuste, Aitor Ibarra (Auditorium)		
2:50pm	Tutorial Session B • "Licensing and Property Rights" with Alberto Accomazzi, Kelly Lockhart, & Jan Reenrick (Café Con) • "Tokenization and Context Engineering" with John Wu (Auditorium)		
3:50pm	Coffee Break		Café
4:00pm	Tutorial Session C • "AI and Trust" with Jonathan Hargis (Café Con) • "Retrieval-Augmented Generation with Sultan Hassan (Auditorium)		
5:00pm	**Welcome Reception (5:00pm-7:00pm)**		**Café**

Shuttles will be available back to the host hotels.

Tuesday, March 10, 2026

Time	Title	Speaker	Location
8:30am	Breakfast (Café) **Speaker Tech Check for Session 1 (Auditorium)**		
Session 1 - Chair: Kelly Lockhart, Moderator: John Soltis			
9:00am	Welcome	Jenn Lotz, STScI Director Michelle Ntampaka, Chair	Auditorium
9:20am	Ethics & Social Sciences at the Intersection of NLP	Anjalie Field (I)	Auditorium
9:50am	Toward Community Standards for Generative AI in Astronomy	Ivelina Momcheva (C)	Auditorium
10:00am	Certifying the AI-Generated Stories We Tell About the Cosmos	Shanil Virani (C)	Auditorium
10:10am	Learning to Reason About Astronomy-Specific Technical Tasks with Large Language Models	Sanjib Sharma (C)	Auditorium
10:20am	Summary and Next Steps	Michelle Ntampaka, Chair	Auditorium
10:30am	**Poster Pops:** Adele Plunkett, Simone Astarita, Nicholas Susemiehl, Sarah Burke-Spolaor		
10:45am	Break (Café) **Speaker Tech Check for Session 2 (Auditorium)** **Posters Displayed:** Adele Plunkett, Simone Astarita, Nicholas Susemiehl, Sarah Burke-Spolaor		
Session 2 - Chair: Jan Reerink, Moderator: John Soltis			
11:15am	AI tools to enable literature research	Kartheik Iyer (C)	Auditorium
11:45am	Lessons from Nancy: Open RAG Infrastructure for Research Agents	Amber Malpas (C)	Auditorium
11:55am	From Authorship to Expertise: Scalable Name Disambiguation and Reviewer Matching	Vincente Amado Olivo (C)	Auditorium
12:05pm	Multilabel Text Classification for Concept Assignment in Astrophysics Literature	Atilla Alkan (C)	Auditorium
12:15pm	AstroGenesis: A Domain-Aware Multi-Agent Model for Data-Driven Astrophysics	Ashish Mahabal (C)	Auditorium
12:25pm	Summary and Next Steps	Michelle Ntampaka, Chair	Auditorium
12:35pm	**Poster Pops:** David Rodriguez, Paco Holguin, Josh Speagle & Biprateep Dey, Jose Antonio Najera		
12:50pm	Lunch (Café)		
Session 3: Hack Time			
1:50pm	Breakout Discussion Session	Auditorium, Café Con, Café	
2:25pm	Hack Group Pitches		Auditorium
3:25pm	Break (Café) **Posters Displayed:** David Rodriguez, Paco Holguin, Josh Speagle & Biprateep Dey, Jose Antonio Najera		
3:55pm	Hack Group Work Time	Auditorium, Café Con, Café	
5:00pm	End of Day		Auditorium
5:30pm	**"Un-Banquet" at R.House *Optional*** 301 W. 29th St. Baltimore, MD 21211 (5:30pm-7:30pm)		

Shuttles will be available back to the host hotels. There is no shuttle service to or from R. House.

Wednesday, March 11, 2026

Time	Title	Speaker	Location
8:30am	Breakfast (Café) **Speaker Tech Check for Session 4 (Auditorium)**		
Session 4 - Chair: Kelly Lockhart, Moderator: Ana Maria Delgado			
9:05am	Automated Mission Classification with Large Language Models	John Wu (I)	Auditorium
9:35am	Language AI-Driven Automation of MAST Bibliographic Classification	Jinmi Yoon (C)	Auditorium
9:45am	IRSA's prototype API assistant	Michael Jones (C)	Auditorium
9:55am	ESOFinder: an LLM-powered tool to help users navigate ESO documentation	Paula Sanchez Saez (C)	Auditorium
10:05am	Exploring Multi Agent Systems to support archival data discovery in Euclid	Miguel Doctor Yuste (C)	Auditorium
10:15am	Summary and Next Steps	Michelle Ntampaka, Chair	Auditorium
10:25am	**Poster Pops:** Aitor Ibarra, Lorenz Ehrlich, Ryan Thrill, Rafael Martinez-Galarza		
10:40am	Break (Café) **Speaker Tech Check for Session 5 (Auditorium)** **Posters Displayed:** Aitor Ibarra, Lorenz Ehrlich, Ryan Thrill		
Session 5 - Chair: Jan Reerink, Moderator: Ana Maria Delgado			
11:15am	Bringing the arxiv into focus with aparture (with podcasts?!)	Joshua Speagle (I)	Auditorium
11:45am	AION-Search: Semantic searching galaxy images with synthetic captions	Nolan Koblischke (C)	Auditorium
11:55am	From Prompts to Queries: Accessing Astronomical Data using Natural Language	Roman Machacek (C)	Auditorium
12:05pm	A text-to-SQL framework for the SDSS SkyServer database	Peter Zsoldos (C)	Auditorium
12:15pm	Group Photo		Front Step
12:30pm	Lunch (Café)		
Session 6: Hack Time			
1:30pm	Hack Group Work Time	Auditorium, Café Con, Café	
3:00pm	Break (Café) **Posters Displayed:** Rafael Martinez-Galarza, Rohit Raj, Dan Coe		
3:30pm	Hack Group Work Time	Auditorium, Café Con, Café	
5:00pm	**SDAS Event (5:00pm-7:00pm)**		**Café**

Shuttles will be available back to the host hotels.

Thursday, March 12, 2026

Time	Title	Speaker	Location
8:30am	Breakfast (Café) **Speaker Tech Check for Session 7 (Auditorium)**		
Session 7 - Chair: Alberto Accomazzi, Moderator: Mikaeel Yunus			
9:05am	Benchmarking and Evaluating NLP models	Jessy Li (I)	Auditorium
9:35am	Using Language Models to Scale Expert Morphology Classifications of Nearby Galaxies	Michael Koss (C)	Auditorium
9:45am	Large Language Model Driven Analysis of General Coordinates Network (GCN) Circulars	Vidushi Sharma (C)	Auditorium
9:55am	**Poster Pops:** Kiera McCormick		
10:10am	Summary and Next Steps	Michelle Ntampaka, Chair	Auditorium
10:20am	Break (Café) **Speaker Tech Check for Session 8 (Auditorium)**		
	Posters Displayed: Kiera McCormick, Ben Ramsey, Simone Riggi		
Session 8 – Hack Group Presentations - Chair: Alberto Accomazzi, Moderator: Mikaeel Yunus			
10:50am	Text isn't all you need	Mike Smith (I)	Auditorium
11:20am	Hack Group Coordination and Prep **Group presentations should be uploaded by 12:20pm**	Auditorium, Café Con, Café	
12:20pm	Lunch		Café
1:20pm	Hack Group Coordination and Prep		
1:50pm	Hack Group Presentations		Auditorium
3:20pm	Coffee Break (Café) **Posters Displayed:** Yang Cheng, Hayley Roberts, Xiaowen Zhang		
3:50pm	Hack Group Presentations		Auditorium
4:40pm	Summary and Next Steps	Michelle Ntampaka, Chair	Auditorium

Shuttles will be available back to the host hotels.

Workshop Abstracts

STScI

Invited Talks

Ethics & social sciences at the intersection of NLP

Anjalie Field
Johns Hopkins University

Evaluation and auditing of AI systems is becoming increasingly difficult as people use systems in a wide variety of ways, with instructions expressed in natural language prompts. We can no longer use readily quantifiable metrics like accuracy or statistical parity to understand model performance and potential impacts. Instead, we need ways of conducting open-ended analyses of models and usage data that do not infringe on user privacy. In this talk, I will discuss ways we are working towards these goals, beginning with an in-depth analysis of LLM usage in a specific domain: AI for querying astronomy literature. While manual analysis of usage data and follow-up interviews with astronomers offer an in-depth look at how astronomers interacted with an LLM-powered system, manual evaluation does not scale to the large volume of usage data in other contexts. Thus, I will next discuss methods for automated inductive coding, which offer more scalability, and finally, I will very briefly discuss text anonymization approaches to increase oversight of model usage without compromising privacy.

Your Pathfinder research

Kartheik Iyer
Flatiron Institute (CCA)

Text isn't all you need

Mike Smith
Max Planck Institute for Astronomy

We will explore how foundation models are expanding beyond text to transform astronomical research. We'll discuss two key architectural approaches to multimodality: late fusion models (like LLaVA) that combine separate encoders for different data types, and early fusion models that enable multimodal interaction from the ground up. As a case study in late fusion, we'll examine AstroLLaVA -- a vision-language model fine-tuned on ~30k astronomical image-text pairs that can engage in natural dialogue about cosmic phenomena. For early fusion, we'll look at AstroPT, a GPT-style autoregressive transformer trained on 8.6M galaxy images that demonstrates predictable scaling laws similar to language models, with downstream task performance improving systematically with model size. Finally, we will explore new results from the Platonic Universe project, which tests whether different foundation models trained on fundamentally different data are converging toward shared representations of underlying astrophysical reality. Our findings suggest they are: larger models show increasing representational alignment regardless of architecture or training modality. This implies that -- rather than building astronomy-specific architectures from scratch like AstroPT -- we may be better off in practice if we adapt existing pre-trained models (like the models used in AstroLLaVA).

AI for Scientists: Learning and Doing in the Era of VLMs

Joshua Speagle
University of Toronto

While "Chat" and "AI" have become ubiquitous, their impact on academia (from learning to research and everything in between) has been harder to quantify. Nevertheless, there remains a pronounced gap between the promise of AI and its actual impact among researchers. Using data from the adoption of MCMC and AI/ML in astronomy, I will highlight how human-centric design targeting practitioners continues to play an outsized role in adoption but not in increased expertise. I will then highlight how these practices have impacted widespread adoption of these technologies, but how their default use often hampers skill acquisition (with examples in coding). I'll then look at how we can use "backwards design" principles to combine these two findings into a framework that can guide judicious usage and development of AI in astronomy, with an eye towards many of the topics discussed at the workshop. I'll try to leave plenty of time for open discussion.

Automated Mission Classification with Large Language Models

John Wu
Space Telescope Science Institute

Telescope bibliographies record the pulse of astronomy research by capturing publication statistics and citation metrics for telescope facilities. However, the growing rate of publications threatens to outpace our ability to manually label astronomical literature, and jeopardizes our ability to measure the scientific impact of astronomical facilities and archives. I will present the Automated Mission Classifier (AMC), a tool that uses large language models (LLMs) to identify and categorize telescope references by processing large quantities of paper text. I will describe the multi-stage LLM system design that drives strong performance on real-world bibliographic data; AMC achieves ~0.95 F1 score on classifying astronomy preprints as JWST science papers, and a modified version of AMC placed third (F1 score of 0.84) in the TRACS Shared Task Kaggle competition for classifying HST, JWST, and Chandra papers. I will also discuss how AMC can be used to interrogate historical datasets and surface potential label errors. This work demonstrates that LLM-based applications offer powerful and scalable assistance for library sciences.

Contributed Talks

Multilabel Text Classification for Concept Assignment in Astrophysics Literature

Atilla Alkan
Center for Astrophysics | Harvard & Smithsonian

Automatically tagging scientific articles with controlled vocabulary concepts is essential for indexing, retrieval, and knowledge organization, especially in domains like astrophysics, where the volume of published papers continues to grow rapidly. In this ongoing work, we investigate the problem of concept assignment, where papers can be tagged with multiple concepts from a large and highly imbalanced label space. Building on a corpus of 21,702 astrophysics abstracts from NASA ADS annotated with 1,860 concepts from the Unified Astronomy Thesaurus (UAT), we conduct a systematic exploration of computational approaches to concept classification. Our study spans rule-based baselines, nearest-neighbor retrieval using text embeddings, and fine-tuning pretrained language models for multilabel prediction. Building on these experiments, we then focus on transformer-based models and examine post-hoc methods aimed at improving their prediction quality. For instance, a first approach uses pointwise mutual information (PMI) between co-occurring concepts to recalibrate label probabilities, yielding a modest gain of roughly +1 F1 point. A second approach leverages a large language model (LLM) as a reranker that re-evaluates candidate concepts, showing promising potential (+5 F1 point). These findings remain preliminary and are expected to evolve as we refine our models. We also experiment with evaluation methods that incorporate relationships between concepts within the UAT's hierarchical structure, enabling us to better assess the relevance of model-predicted concepts, including those not explicitly present in the ground-truth annotations but semantically close to them. In perspective, our aim is to provide a clearer picture of the opportunities and challenges in large-scale concept assignment for specialized scientific domains, and provide tools that could assist librarians, data curators, and researchers in astrophysics.

From Authorship to Expertise: Scalable Name Disambiguation and Reviewer Matching

Vicente Amado Olivo
Michigan State University

The exponential growth of research outputs (e.g., papers and proposals) is overwhelming our traditional research workflows from literature searches to peer review. As the volume of research expands, finding the right experts to evaluate, connect collaborators, or support large-scale scientific teams has become increasingly difficult. We present scalable techniques for author name disambiguation and expertise representation to identify experts worldwide across large-scale literature with metadata constraints. We developed the Neural Author Name Disambiguator to construct the accurate publication histories

required for robust expertise modeling. This Siamese neural network addresses the challenge using only widely available metadata—names, titles, and abstracts. Trained on a large ORCID-linked physics and astronomy dataset, the model achieves over 94% accuracy in pairwise disambiguation and 95% F1 in clustering. Leveraging the disambiguated profiles, we systematically evaluated expertise representation, ranging from TF-IDF to Transformers, to optimize performance on a peer reviewer matching task in collaboration with the European Southern Observatory. Our initial evaluation demonstrates that statistical embeddings (e.g., TF-IDF) outperform other methods in identifying relevant experts. By integrating author disambiguation with expertise representation, we move beyond tracking authorship to connecting the right expertise to where it's needed most.

Exploring Multi Agent Systems to support archival data discovery in Euclid

Miguel Doctor Yuste
European Space Agency – ESAC

AI is progressing very rapidly and its application to different fields of science, including astronomy, is producing remarkable results. Large Language Models (LLMs), with their capabilities in text comprehension and generation, are reshaping how researchers interact with complex digital platforms. Building on this foundation, Multi-Agent Systems (MAS) further extend AI's potential by enabling multiple intelligent agents to collaborate within shared environments to achieve common objectives or approach intricate tasks. The ESA Euclid Mission is surveying one third of the sky with unprecedented resolution generating an estimated 25–30 petabytes of data products that will be made available through the Euclid Data Space on the Euclid Science Archives and ESA Datalabs Science Platform. For the particular case of the higher level science data products (also known as Level 3 products), over 200 different types of products will be generated by the Euclid Consortium and released on Euclid Data Space. Keeping the users up to date with documentation, formats, and availability of data is challenging. In this context, we explore how MAS can facilitate the discovery and use of scientific products. Namely, we implement a proof-of-concept MAS orchestration tool designed to automate scientific workflows and simplify user access to Euclid data. Through natural language interaction, the tool enables deeper exploration of the Euclid Data Model and documentation, generates ADQL queries aligned with user interests, and produces executable Jupyter notebooks within the ESA Datalabs platform.

IRSA's prototype API assistant

Michael Jones
Caltech/IPAC

The Infrared Science Archive (IRSA) hosts an extensive collection of astronomical datasets, along with powerful search, visualization, and programmatic interfaces. While expert users can effectively leverage these capabilities, many other users can struggle to navigate and parse the documentation, understand mission-specific data structures, or construct complex queries. LLMs present an opportunity to lower this barrier to entry, while still maintaining the depth of the existing interfaces. To this end, we are developing a prototype LLM-based query assistant built on retrieval-augmented generation (RAG) and OpenAI models. This prototype leverages IRSA documentation, schema metadata, and curated examples, to enable users to ask natural-language questions such as "Which datasets cover

this sky position?" or "How do I retrieve all WISE sources brighter than X?" The assistant generates grounded explanations, constructs the relevant archive queries (e.g., ADQL, or API calls), and guides users through refinement steps.In this presentation, we will discuss how well the prototype performs in practice, including generated query accuracy, common failure modes, and operational considerations. We will also cover lessons learned on topics such as, how to improve retrieval quality and how to strengthen guardrails against hallucinations.

From Prompts to Queries: Accessing Astronomical Data using Natural Language

Roman Machacek
University of Bern

Accessing and querying astronomical data can be challenging, especially for early-career scientists, due to the technical demands of the Astronomical Data Query Language (ADQL) and the complexity of related documentation. While big AI assistants, such as ChatGPT, can aid in translating natural language to ADQL, they are resource-intensive and may not be practical for all users. In this work, we investigate the potential of small language models (SLMs), for generating ADQL queries from natural language prompts. Starting with50GB of raw Gaia DR3 queries, we curated a dataset of 30,000 samples and enriched it with multiple natural language descriptions using LLMs. We further utilize LLMs to annotate the dataset with complexity of the query along with its description. We evaluate newly created dataset with several approaches, including zero shot, few-shot, and reasoning-based inference. To provide richer context, we use Gaia DR3 documentation and implement retrieval-augmented generation (RAG), enabling smaller models to dynamically access domain-specific knowledge. Additionally, we explore low-rank adaptation (LoRA) fine-tuning to further push performance of the models. Our results indicate that pre-trained small LLMs, especially when combined with RAG and LoRA, can help users in generating ADQL queries from natural language, offering a practical alternative to larger AI assistants. This approach lowers technical barriers, supports astronomers in accessing and analyzing data, and demonstrates the potential for lightweight, domain-specialized models in astronomical research.

AstroGenesis: A Domain-Aware Multi-Agent Model for Data-Driven Astrophysics

Ashish Mahabal
California Institute of Technology

Astrophysics is entering a data-rich era driven by multi-wavelength observatories and multi-messenger experiments. These facilities produce vast, heterogeneous datasets that challenge traditional analysis pipelines. General-purpose AI systems, while powerful, often lack the contextual reasoning and scientific rigor required for astrophysical interpretation. AstroGenesis is an AI-powered, domain-aware multi-agent research assistant designed to revolutionize how astrophysicists access, analyze, and interpret astronomical data. Built upon a Retrieval-Augmented Generation (RAG) framework and a modular multi-agent architecture, AstroGenesis integrates literature retrieval, data access, theoretical modeling, and hypothesis generation into a unified ecosystem. Each specialized agent autonomously handles dedicated tasks-such as spectral fitting, time-series analysis, or model inference-

under the coordination of a central supervisory agent that ensures transparency and reproducibility. Key innovations include a domain-aware RAG system that grounds responses in peer-reviewed literature; seamless integration with multi-wavelength and multi-messenger archives for both raw and processed data; and neural-network-based modeling agents trained on large-scale radiative simulations for real-time, physics-consistent inference. It is possible to inspect the machinery with a human-on-the-loop for validation, and feedback to further enhance reliability. Demonstrated through a prototype for blazar research, AstroGenesis can be generalized to diverse astrophysical phenomena. By unifying reasoning, data retrieval, and theoretical modeling within a scalable framework, it lowers the barrier to advanced analysis and fosters transparent, reproducible, cross-disciplinary discovery in modern astrophysics.

Lessons from Nancy: Open RAG Infrastructure for Research Agents

Amber Malpas
The Ohio State University

Nancy began as a Slack-based support agent for the Roman Microlensing Data Challenge 2026, built on a modular retrieval-augmented generation (RAG) stack that ingests community tools, documentation, notebooks, and literature. Her architecture has since evolved into a reusable infrastructure that other science agents can adopt with minimal overhead. This presentation will outline the system's design (modular ingestion pipelines, adaptive chunking, dual-embedding retrieval, cached hierarchical summaries, and model-in-the-loop refinement) as well as its deployment through Slack, HTTP, and the Model Context Protocol. I will also discuss current challenges, including extracting semantic value from code and PDFs, benchmarking weighting strategies, and supporting efficient incremental updates to large knowledge bases. Nancy's "brain" now serves as a community-maintained retrieval layer for AI research agents, reducing duplication of effort and enabling researchers to build specialized tools without reconstructing the retrieval stack for every new application.

Toward Community Standards for Generative AI in Astronomy

Ivelina Momcheva
Max Planck Institute for Astronomy

As the adoption of generative AI accelerates across astronomy, from coding to paper writing to data analysis, our community urgently needs practical guidance on how to use these tools responsibly. However, there is a significant disagreement on the appropriate community norms and perhaps there are variations depending on tasks and context. In this talk, I present a set of ongoing initiatives aimed at building such shared norms and actionable recommendations for ethical use of generative AI within astronomy research environments. The efforts described include work to develop institute-level guidelines for the use of GenAI and policy work within the Journal of Open Source Software (JOSS), where we are formulating editorial standards for the evaluation of AI-assisted code, among others. I will also draw on outcomes from several interactive workshops I have led on

ethical AI, including a participatory game, The Grayzone of Ethical AI, that helps researchers confront ambiguous, real-world situations where “best practice” is not obvious. As part of the talk I hope to also engage the audience in a discussion of appropriate guidelines around GenAI. Together, I hope we can chart a path for developing community-driven guidelines that empower astronomers to use generative AI with confidence, creativity, and integrity. Optionally, I can offer to run a session of the Grayzone of Ethical AI or have the workshop participants generate a new version of the game based on their experiences.

ESOFinder: an LLM-powered tool to help users navigate ESO documentation

Paula Sanchez Saez
European Southern Observatory – Germany

The large amount and diversity of documentation available for users of the European Southern Observatory (ESO) — including Phase 1, Phase 2, and Phase 3 guidelines, instrument manuals, data-reduction pipeline documentation, and archive interfaces — makes it increasingly challenging for the astronomical community to efficiently find relevant information. To address this, we have developed ESOFinder, an in-house chatbot powered by Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). ESOFinder integrates public information from instrument manuals, phase 1/2/3 documentation, data reduction pipeline manuals, the ESO Knowledge Base, and key web resources to provide concise, context-aware, and reference-linked answers to user queries about proposal/observation preparation, data retrieval, and data processing. Built on open-source European LLMs (Mistral AI models), ESOFinder ensures data privacy, transparency, and complete control over the knowledge base. Its multi-step architecture allows verification of retrieved documents and generated answers, reducing the risk of hallucinations and improving the reliability of responses compared to generic tools like ChatGPT or Gemini. The current version of ESOFinder is being tested internally at ESO to evaluate its performance, assess its integration with internal workflows, and identify limitations in coverage and accuracy. These tests will guide further improvements, including the incorporation of additional documentation and enhanced retrieval strategies. Ultimately, ESOFinder aims to become an interface for users to navigate ESO’s complex observing documentation ecosystem and to support both staff and community astronomers in their daily tasks. In this talk, I will introduce ESOFinder, describe its design—including the RAG pipeline and answer-generation strategy—and present the quality assessment of the tool based on feedback collected from ESO staff.

Learning to Reason About Astronomy-Specific Technical Tasks with Large Language Models

Sanjib Sharma
Space Telescope Science Institute

The Large Language Models (LLMs) have now started showing promise in domain-specific reasoning tasks. This enables them to be used as AI-agents that can speed up science research. The key to improving such models is to apply reinforcement learning (RL) on benchmarks with verifiable facts which, unlike maths and coding, is challenging for subjects like astronomy. However areas such as technical documentation of astronomy observatories and their data sets allow for the creation of domain-specific, verifiable benchmarks. We use technical and instrument reports related to the Hubble, JWST and Roman Space Telescopes to develop question (task) and answer (solution) pairs that test domain-specific reasoning. This benchmark is applied to evaluate various existing LLMs.

Large Language Model Driven Analysis of General Coordinates Network (GCN) Circulars

Vidushi Sharma
University of Maryland Baltimore County

The General Coordinates Network (GCN) is NASA's time-domain and multi-messenger alert system. GCN distributes two data products - automated ``Notices,'' and human-generated ``Circulars,'' that report the observations of high-energy and multi-messenger astronomical transients. The flexible and non-structured format of GCN Circulars, comprising of more than 40500 Circulars accumulated over three decades, makes it challenging to manually extract observational information, such as redshift or observed wavebands. In this work, we employ large language models (LLMs) to facilitate the automated parsing of transient reports. We develop a neural topic modeling pipeline with open-source tools for the automatic clustering and summarization of astrophysical topics in the Circulars database. Using neural topic modeling and contrastive fine-tuning, we classify Circulars based on their observation wavebands and messengers. Additionally, we separate gravitational wave (GW) event clusters and their electromagnetic (EM) counterparts from the Circulars database. Finally, using the open-source \texttt{Mistral} model, we implement a system to automatically extract gamma-ray burst (GRB) redshift information from the Circulars archive, without the need for any training. Evaluation against the manually curated \textit{Neil Gehrels Swift Observatory} GRB table shows that our simple system, with the help of prompt-tuning, output parsing, and retrieval augmented generation (RAG), can achieve an accuracy of 97.2\% for redshift-containing Circulars. Our neural search enhanced RAG pipeline accurately retrieved 96.8\% of redshift circulars from the manually curated database. Our study demonstrates the potential of LLMs, to automate and enhance astronomical text mining, and provides a foundation work for future advances in transient alert analysis.

Certifying the AI-Generated Stories We Tell About the Cosmos

Shanil Virani
Astronomical Society of the Pacific

Large language models are rapidly transforming how astronomy stories are written, narrated, and visually composed, yet their fluency often obscures a central concern: distinguishing explanations that are scientifically grounded from those that only sound authoritative. In response, the Astronomical Society of the Pacific has developed the Generative AI Astronomy Video Certification Program, a workflow for assessing the scientific integrity of AI-generated astronomy videos and the language-based assertions within them. The initiative treats these videos not just as media artifacts but as narrative acts about the cosmos, carrying authority, shaping public understanding, and requiring structured mechanisms of trust. The certification process blends human scientific review with novel verification infrastructure. Approved videos receive a digital hologram certification mark for short-form platforms, and each certification event is recorded on a public Web3 ledger, providing a tamper-resistant provenance chain and durable transparency in how science claims are validated. The evaluation rubric centers on recurring dimensions of LLM-generated content surfaced during pilot reviews: scientific accuracy, source integrity, clarity, ethical and cultural responsibility, and visual integrity. This structure enables a consistent, human-in-the-loop certification process that makes evaluation standards transparent to creators, guiding them toward more accurate astronomy storytelling. For the Language AI in the Space Sciences workshop, we present this framework as a model for responsible deployment of language technologies in astronomy communication. We describe how the program supports shared evaluation-process standards—transparent, repeatable methods for checking scientific claims, identifying common error patterns, and recording provenance so creators and audiences can trust when AI-generated astronomy content has been properly reviewed. This builds trust in astronomy content.

Language AI-Driven Automation of MAST Bibliographic Classification

Jinmi Yoon
Space Telescope Science Institute

We present early results from the Bibliography Classification Automation Tool (BibCAT), a Python based system that applies language AI methods to measure bibliometrics for the Mikulski Archive for Space Telescopes (MAST). BibCAT automates the classification of candidate publications into “science” and “non-science” categories for each MAST mission based on whether the paper makes use of MAST data. This approach provides a scalable way to maintain consistent, mission level publication metrics as the volume of astronomical literature continues to grow. MAST at the Space Telescope Science Institute, hosts data collections from more than twenty NASA missions and ground-based observatories spanning over fifty years. These datasets have contributed to scientific discoveries that have shaped our understanding of the Universe, as reflected in the

astronomical literature. Bibliometrics provide a quantitative means to measure both the productivity of hosted missions and their contribution to astronomical research. However, a bibliometric study requires identifying mission relevant publications and determining whether each paper uses MAST data for scientific analysis ("science") or does not ("non-science"). This task has traditionally relied on extensive manual review, limiting consistency and scalability. As the number of MAST related publications increases, especially with upcoming data intensive missions like the Roman Space Telescope, this manual approach is no longer sustainable. BibCAT can reduce manual workload, improve classification consistency, and enable more timely updates to publication metrics. In this presentation, we highlight BibCAT's goals, system design, and initial performance, demonstrating how language AI methods can enhance bibliometric tracking and support the evaluation of scientific impact across the space sciences. We also welcome discussion on limitations, failure modes, and opportunities for future improvement.

A text-to-SQL framework for the SDSS SkyServer database

Peter Zsoldos
Mid Sweden University

The Sloan Digital Sky Survey (SDSS) offers one of the largest and most influential astronomical datasets, yet accessing it often requires writing complex SQL queries - a challenge that limits its reach beyond expert users. Recent advances in large language models (LLMs) open new possibilities for interacting with structured databases through natural language, making scientific data more approachable and enabling new forms of education, discovery, and collaboration. We present an LLM-powered Retrieval-Augmented Generation (RAG) system that allows users to query the SDSS database simply by asking questions in plain English. Built on the Vanna framework, the system combines semantic vector search with a language model to translate natural language requests into executable SQL queries. A custom knowledge base (KB) grounded in SDSS documentation and schema enhances both precision and contextual relevance. Experimental evaluations across several LLMs demonstrate that RAG-based retrieval substantially improves SQL generation accuracy compared to baseline models without schema context. Notably, the LLaMA3-1.8Bmodel achieved the highest relative gain, with a six-fold improvement in pass@1 accuracy when using the KB. The framework enhanced the performance by enforcing SDSS-specific SQL idioms and reducing syntax errors. Surprisingly, from all the models evaluated, the reasoning ones performed poorly in comparison to smaller non-reasoning models, thus highlighting both the power of small models, which are adapting better to external context and the failure of the system to eliminate reasoning errors. By lowering barriers to data exploration, this work invites a wider community of researchers, students, and citizen scientists to engage directly with astronomical data. We aim to spark discussion on how natural language interfaces can support hands-on experimentation, education, and collaboration across the space sciences.

Posters

Automatic Construction of an Extragalactic Distance Catalogue

Jose Antonio de Jesus Najera Quintana
University of Portsmouth

Redshift-independent distance measurements underpin most of astrophysics. There are a handful of distance catalogues in the literature, like the NASA/IPAC Extragalactic Database of Distances (NED-D) and Cosmic Flows 4 (CF4). The purpose of this project is to update them and check for potential systematics in the distance ladder. To achieve this we are developing a code to extract distances from literature studies in an automatic way by using regular expressions (regex) and large language models (LLMs). We make use of the Kaggle arXiv dataset, storing all astro-ph articles on our local cluster ready for processing. To ensure the code has broad coverage, we iteratively construct and refine regular expressions by manually analyzing a small number of arXiv papers until linguistic variation plateaus. We handle cases where the distance modulus is reported along with the uncertainties both in text and tables, and will pass any ambiguous or potential cases to an LLM to verify. Once the code is reliably extracting distances from these papers, we will use our local cluster to run the code over all arXiv papers in the astro-ph category. To validate the results, we test whether we recover the same papers and distances as NED-D, with a goal of a comprehensive and accurate dataset to succeed NED-D with updates coming from more recent papers. This will allow us to have a reliable extragalactic-distance catalogue useful for precision and accuracy cosmology and reconstructions of the cosmic distance ladder.Co-Authors: Varsha P. Kulkarni, Viacheslav V. Klimenko

How We (Co-)Write Astronomy with LLMs

Simone Astarita
Space Telescope Science Institute

The ease of access to and the capabilities of large language models (LLMs) are changing scientific research, and astronomy is no exception. As space-specific tools proliferate (e.g., Pathfinder, AstroCoder, AstroAgents) and models expand their knowledge of the field, interest in studying the effects of AI on astronomy research and paper writing grows. I briefly survey previous work on how LLMs might have spiked the use of certain words, highlighting the limitations of the methods and data used so far. I then introduce a dataset of over 300,000 astronomy abstracts compared with two AI versions, one produced by ChatGPT 3.5 and one by 4o, the largest AI-vs-human dataset on a specific subject and task. First, I discuss how I crafted this dataset to reflect the actual usage of ChatGPT as best as possible. I then briefly illustrate its potential to differentiate between human- and AI-produced short text, but explain why this problem remains difficult. Finally, I present some promising results on the effects of LLMs on the population level, showcasing statistical techniques to assess how LLMs modify our lexicon and writing in general.

BOBcat: Catalog of Binary Supermassive Black Holes

Sarah Burke-Spolaor
West Virginia University Research Corporation

We are working to create BOBcat, an online tracking system for binary supermassive black hole candidates. In its present form, we will be homogenizing black hole binary orbital parameters that have been published for candidate dual and binary systems up to separations <~1kpc, capturing information that is otherwise difficult to extract from the literature. We are also working on enabling this resource as part of a broader multi-messenger infrastructure that can accept or update automatically submitted candidate parameterizations from systems like ZTF, LSST, or LISA. We are exploring the reliability of language AI options to manage a reliable, long-term maintainable automatic literature identifier and extractor for BOBcat.

Attention-Pooling motivated multi-modal SSL model for versatile imaging + spectra fusion

Yang Cheng
Max Planck Institut for Astronomy

The development of self-supervised (SSL) multimodal models has grown rapidly in NLP and computer vision, but their application to scientific domains remains less well explored. Astronomy provides an ideal testbed for such models due to its large publicly available data and diverse modalities. Recent work has begun exploring such foundation models with promising results in tasks e.g. morphology classification, redshift inference, etc. However, existing multimodal approaches still lack tolerance to missing data and do not provide a principled method for leveraging a shared latent space. Current designs rely on: (i) early fusion of raw modalities (cannot tolerate missing data/masking); (ii) late fusion via embedding concatenation (is suboptimal and redundant); or (iii) single-modality downstreamtasks (AstroCLIP). These limits make them less suitable for realistic survey data, where not all imaging or spectral bands are available.Therefore, we introduce a multi-modal SSL model using NIRISS/NIRCam imaging and spectra, built explicitly for missing-data tolerance and late embedding fusion through attention pooling. Each individual imaging band and each spectrum is treated as its own modality rather than merging bands into a single encoder. For images, we adopt off-the-shelf SSL backbones including DINOv2 and MoCo v2. For spectra, following AstroCLIP, we pretrain a GPT-2–like transformer tokenizer with self-attention to obtain spectral embeddings. After pretraining, all embeddings are aligned through projection heads optimized with a contrastive loss. Finally, we apply an attention-pooling module—like mechanisms used in image-text models—where a small network assigns SoftMax-normalized weights to the input embedding vectors. The fused vector, as a weighted sum, automatically emphasizes informative modalities while assigning near-zero weight to missing inputs, yielding a unified and robust representation for downstream tasks.

AI-assisted science and technical support for astronomers studying the first galaxies with JWST

Dan Coe
Space Telescope Science Institute

For the past two years, I've used AI LLMs extensively in both aspects of my job as an astronomer: doing science and supporting other astronomers using JWST. My science team studies the first galaxies with JWST, and we use AI to help develop code, learn complex topics, and discuss proposal ideas. AI lowers the barrier to entry for astronomy students, offering personalized tutoring and coding assistance, addressing what had previously been the biggest time sink: debugging code. In my role supporting other astronomers using JWST, I've shipped software and calibrations aided by AI. I'll briefly present examples, lessons learned, and best practices. Experience builds both trust in capabilities and understanding of limitations, as I'll discuss based on my thousands of conversations and coding sessions with AI. Having co-led a bi-weekly AI discussion group, I look forward to the discussion at this meeting. Has AI accelerated your work? I estimate AI accelerates 1/4 of mine, though I quickly fill that extra time with new projects I couldn't do before. How should we shape the future? Other fields like medicine and chemistry are beginning to use AI to perform deep research and develop innovative solutions to complex problems. While some astronomy research is fundamentally limited by data, we should identify the open questions about our universe that can benefit most from AI-accelerated discovery.

XMMGPT: Integrating Agentic RAG and Autonomous Agents for Astronomical Data Processing

Lorenz Ehrlich
ESAC – Telespazio

XMMGPT is a dual-purpose project which aims to serve as a unique access point to aid astronomers in their research with XMM-Newton data, and as an exploration of language models and AI systems within European Space Agency (ESA) workflows. The system is comprised of 4 main parts, a heavily customized Agentic Retrieval Augmented Generation (RAG) pipeline, a visibility checker tool, a long-term light curve generator, and an autonomous agent to execute data processing tasks from the Scientific Analysis System (SAS). The RAG system is built from SAS technical documentation with contextual embeddings, naive knowledge graphs, a hybrid fine-tuned text and vector similarity search, as well as an agent that routes user queries to appropriate documentation types. The visibility checker tool transforms natural language queries into API calls to a visibility server which implements the IVOA ObjObsSAP protocol while the long-term light curve generator transforms natural language queries into API calls to gather flux data and generate a representative plot.The autonomous SAS Agent takes natural language queries and goes through a series of sequential steps in order to send API calls to RISA (Remote Interface to SAS Analysis) or ULS (Upper Limit Server) and retrieve valid scientific products.Data privacy and cost concerns have led to the project being developed on relatively small local hardware and a constant challenge has been to find state-of-the-art, smaller footprint, language models which give good performance.Approaches to

autonomous SAS code creation and execution are currently being explored while the project is steadily being updated to take advantage of evolving industry standards.

Enabling high-volume data processing of JWST exoplanet data via generative AI compression

Paco Holguin
The Johns Hopkins University Applied Physics Laboratory

As space science observatories continue to deliver increasing volumes of data, on-board compression may be critical for maximizing scientific return, given limited downlink and onboard storage resources. Even assuming improved bandwidth and data recorders, space missions will still be limited from their full data collection potential. Generative AI models can learn complex, non-linear relationships across high-dimensional data, and those learned models can be employed for efficient data compression. We present an experiment demonstrating generative AI-based compression using large language models (LLMs) and variational autoencoders (VAEs). We implement lossless LLM-based and lossy VAE-based compression techniques using 4D spectra data cubes from James Webb Space Telescope (JWST) exoplanet transit observations as a standardized test case. For lossless compression, we re-train a Llama3 architecture to serve as a strong next-token probability model within arithmetic encoding scheme, driving the compression to the entropy limits. We characterize the trade space among compression ratio, reconstruction error, and computational complexity. We set a baseline of 5x lossless compression via the LLM-enabled method with approximately 1.5 billion parameters. We further analyze the impact of lossy compression on downstream scientific interpretation by running decompressed data through the JWST reduction pipeline, demonstrating how reconstruction errors affect faint signal extraction. Our work highlights potential of generative AI compression to enhance data efficiency in future space science missions.

Agent-Ready "Mini-SAS" Microservices for Language-Driven X-ray Data Analysis

Aitor Ibarra
ESAC

We present Mini-SAS, a suite of purpose built, REST-API micro-applications that wrap canonical XMM-Newton SAS tasks into agent friendly LLM tools, enabling language driven for reliable X-ray analysis. Each Mini-SAS targets one well scoped workflow, for example, image, spectra or lightcurve generation for all XMM-Newton instruments, with explicit inputs/outputs, and determinism. The services advertise machine readable schemas to support autonomous tool use by AI agents, and they return structured results (metadata + science products) suitable for downstream analysis. Beyond convenience wrappers, Mini-SAS integrates with XMMGPT via stable REST endpoints and back-ends for remote execution, its design enforces reproducibility and safety. We outline an evaluation plan spanning: (1) fidelity vs. reference SAS pipelines, (2) agent task-success and latency under realistic prompts, (3) end-to-end user studies. Finally, we sketch a roadmap to a mission agnostic X-ray AI framework (XMM-Newton → NewAthena), standardizing micro-services for spectra, timing, imaging, and source detection so language agents can compose cross-

mission analyses while preserving scientific rigor. Mini-SAS thus bridges expert pipelines and autonomous language agents, turning natural-language intent into auditable, high-quality X-ray results.

Augmenting X-ray Astronomical Representations with Scientific Knowledge through Contrastive Learning

Rafael Martinez-Galarza
AstroAI / Center for Astrophysics | Harvard & Smithsonian

Astronomers have produced large multimodal datasets that include images, spectra, and time series, and that encode physical information about the observed objects. In addition, a large amount of physics-specific knowledge about these objects has been accumulated in the astronomical literature. We introduce a physics-informed representation alignment framework that matches X-ray observations of astrophysical objects and text summaries describing the physical properties of those sources. We perform contrastive learning between data representations learned using X-ray-specific encoders that are aware of the Poisson nature of the data, and LLM representations of text summaries obtained from literature curated by NASA's Astrophysics Data System. We demonstrate the generalization capabilities of the system and evaluate the performance of the post-alignment shared representations for regression tasks. Our contrastive learning framework achieves 20% recall at 1%, demonstrating that meaningful alignment between these modalities is possible and potentially accelerating the interpretation of rare or poorly understood sources. More importantly, by combining X-ray spectral and textual information, we improve on zero-shot regression tasks, with the mean absolute error (MAE) metric of key physical variables increasing by ~20% compared to using X-ray data alone. Using a Mixture of Experts architecture that leverages both unimodal and shared representations yields the best performance. Finally, the outlier analysis on the multimodal latent space reveals objects worthy of follow-up investigation, such as a candidate pulsating ULX (PULX) and an X-ray gravitational lens system.

Encoding and Understanding Astrophysical Information in Large Language Model-Generated Summaries

Kiera McCormick
The Johns Hopkins University

Large Language Models have demonstrated the ability to generalize well at many levels across domains, modalities, and even shown in-context learning capabilities. This enables research questions regarding how they can be used to encode physical information that is usually only available from scientific measurements, and loosely encoded in textual descriptions. Using astrophysics as a test bed, we investigate if LLM embeddings can codify physical summary statistics that are obtained from scientific measurements through two main questions: 1) Does prompting play a role on how those quantities are codified by the LLM? and 2) What aspects of language are most important in encoding the physics

represented by the measurement? We investigate this using sparse autoencoders that extract interpretable features from the text.

Advancing Radio Astronomy with Multimodal AI: Tools for the ALMA Science Archive

Adele Plunkett
National Radio Astronomy Observatory

The Atacama Large Millimeter/submillimeter Array (ALMA) Science Archive is a vast repository of ~2 PB of heterogeneous astronomical data, including spectral data cubes, previews, quality reports, and extensive technical metadata. Maximizing the scientific return from this archive requires tools that can seamlessly bridge the gap between human scientific objectives (natural language) and the complex, diverse data formats (image, tabular, coding). This work presents a sampling of multimodal AI applications that our group has designed to streamline several steps of the ALMA research lifecycle. These include a multimodal LLM for data retrieval, a context-aware coding assistant for programmatic archival access, and a tool for recommending proposal technical setups using NLP on historical project metadata. By integrating disparate modalities—natural language, structured metadata, code and spectral data cubes—these tools aim to lower the barrier to entry for new and existing users, enhance data accessibility, and provide an essential framework for accelerating fundamental astronomical discovery via observatory data archives.

A Source-Grounded AI Workflow for Accessible Astronomy Outreach

Rohit Raj
Smithsonian Institution Astrophysical Observatory

Shadow the Scientists (StS), a University of California Santa Cruz program, provides virtual learning opportunities for students and the general public across the globe to engage via Zoom with researchers across the globe while they are conducting research. The knowledge gap is a challenge though, as complex topics and specialized jargon can limit audience engagement and the depth of their questions.To address this, we are developing a practical workflow using Google's NotebookLM (NLM), a Language AI tool designed for source-grounded reasoning. Unlike general-purpose models, NLM bases its output exclusively on user-provided documents. This grounding is critical for scientific accuracy, as it prevents hallucination by restricting the AI's knowledge base to expert-vetted materials: the StS host researchers' conversation and research materials.For each StS session, we curate a few source documents and use NLM to distill core concepts, define key terms in plain language, and generate a short, conversational script. The tool's Audio Overview feature then converts this script into a 5–8 minute, two-host podcast. An audio primer created from research materials can be played for the audience right before a live StS session. More commonly, participants will listen to an NLM podcast based on the recording of an StS session and published research materials before watching the recording.These primers constitute an effective on-ramp for the audience, creating a shared foundation of knowledge. We find that this leads to more confident and insightful questions from the audience. This workflow provides a high-impact, low-barrier method for researchers—who are generally not AI specialists—to repurpose complex research into

accessible outreach content rapidly. We present this as a creative, hands-on application of Language AI that directly lowers barriers to public engagement with space science, and are exploring its expansion into video primers and a multi-language repository.

StarGateVR AI: A Voice-Driven Multi-Wavelength Data Acquisition and Exploration Tool in VR

Ben Ramsey
Rochester Institute of Technology

Recent astronomical missions are generating an unprecedented influx of new data, and new missions like the Vera C. Rubin Observatory and the Nancy Grace Roman Space Telescope will only continue to compound this tidal wave of new data. Researchers must continue to evolve their methodology to contend with the scope of data available and devise methods to interact with it in efficient and effective ways, utilizing new tools and emergent technologies. The ESA Gaia Space Astrometry mission is a prime example; Gaia is delivering precise positional, kinematic, and photometric measurements for almost two billion sources. We present our work on integrating commercially available artificial intelligence (AI) tools into StarGateVR, a virtual reality (VR) program we are presently developing for the exploration, selection, and analysis of Gaia data. StarGateVR AI is outfitted with a speech-based interface to OpenAI's ChatGPT-4o-mini and GPT-4o, enabling rapid, targeted access to external astronomical databases during StarGateVR Gaia data exploration sessions. We demonstrate the nascent tool's efficacy through examples of studies of nearby stellar groups, applying its AI functionality to augment Gaia DR3 data with observations from other missions spanning the electromagnetic spectrum.

Radio-llava: Advancing Vision-Language Models for Radio Astronomical Source Analysis

Simone Riggi
INAF - Istituto Nazionale di Astrofisica

Vision-language models (VLMs) have recently shown promise in general-purpose reasoning tasks, yet their applicability to domain-specific scientific workflows remains largely unexplored. In this work, we evaluated a series of open-weight and commercial VLMs on six tasks relevant to radio astronomy, including classification of radio sources with different morphology. We also introduced radio-llava, a multi-modal assistant built on the LLaVA architecture and adapted for the radio domain through instruction fine-tuning. In zero-shot mode, commercial models like GPT-4.1 outperform open-weight VLMs on most radio benchmarks. However, radio-llava improves upon both base LLaVA and commercial models across nearly all tasks. Despite these gains, specialized vision-only models still deliver substantially better performance across the board. Additionally, we observed that fine-tuning introduces catastrophic forgetting on general multi-modal tasks, with performance drops up to 40% that can be partly mitigated with a more diverse training dataset or shallow fine-tuning.

Preparing for LLMs in Citizen Science with Zooniverse

Hayley Roberts
University of Minnesota

Zooniverse, the world's largest platform for citizen science with over 2.9 million participants, has a long history of thoughtfully integrating new technologies into volunteer-powered research. As interest in large language models (LLMs) grows across the space sciences, we are taking a considered approach to understanding where these tools may genuinely add value while balancing the needs of both volunteers and researchers. In this poster, we describe our early, exploratory experiments in the context of Galaxy Zoo, where volunteers provide free-text descriptions of galaxies, and we investigate whether an LLM applied to this language data, used alongside our existing foundational galaxy model, Zoobot, might help surface unusual or unexpected objects. Our aim is to identify appropriate use cases, limitations, and design considerations for responsible integration of LLMs within citizen science. Through these early experiments, we hope to encourage discussion about how platforms like Zooniverse can approach emerging LLM methods in ways that preserve accessibility, scientific rigor, and community engagement.

Building an MCP Server for MAST queries

David Rodriguez
Space Telescope Science Institute

We introduce a prototype Model Context Protocol (MCP) server designed to provide access to MAST resources. The server integrates several capabilities, including querying observational data with astroquery and accessing exoplanet information with the exomast API. In this presentation we demonstrate tools we've built, show how you can run them yourself, and share what we learned with this prototype.

A Textbook on AI Drafted by AI

Joshua Speagle
University of Toronto

Writing is hard! And as anyone who's tried out current LLMs can attest, while they are great for editing and modifying, they really struggle with "good writing" from scratch. Starting in Summer 2025, I began work on trying to integrate LLMs directly into the entire research writing workflow as part of a project to write up a pedagogical textbook on deep learning ("A Conceptual Introduction to Deep Learning"). A few months and 100+ pages of text later, I am happy to share lessons learned and experiments tried, all the way from ideation (dictation, recording, brainstorming, etc.) to drafting (deep research, summarization, critiquing, etc.) to writing, editing, and polishing (editing, citations, coding, reproducibility, etc.). This will span all ranges of LLM/AI usage, from simple chatbots to fully agentic workflows. My hope is that some of my experiences will help others develop collaborative, human-centred AI-assisted writing workflows with LLMs and agents.

AI for Data Ingestion into IPAC Archives

Nicholas Susemiehl
IPAC, California Institute of Technology

The data archives at IPAC, including the NASA Extragalactic Database (NED) and NASA Exoplanet Archive (NEA), have served as repositories for data published in the astronomical literature for decades. Throughout this time, extracting data from journal articles has remained a challenging task and future large data releases will exasperate this problem. We seek to accelerate the rate at which data can be extracted from journal articles and reformatted into database load files by leveraging recent advances in natural language processing enabled by AI. We are developing a new suite of tools to semi-automate information retrieval from scientific journal articles. Manual methods to extract and prepare data, which can take hours for some articles, are being replaced with AI-powered tools that can compress the task to minutes. A combination of AI and non-AI methods, along with human supervision, can substantially accelerate archive data ingestion. Challenges remain for improving accuracy, capturing data in external files, and flagging issues such as mislabeled object names and missing metadata.

An AI Astronomer for Automated Interpretation of Chandra X-ray Sources

Ryan Thill
Astromind, Inc.

As we enter the era of Large Astronomical Surveys, astronomers are exploring how LLMs and other AI tools can be incorporated into data analysis and domain-specific interpretation. We present an LLM-based "AI astronomer" that ingest raw Chandra X-ray photon recordings corresponding to specific astrophysical sources and returns a rich, human-interpretable characterization of each source. At its core is a photon event analyst: an LLM fine-tuned on question and answer pairs distilled from two and a half decades of Chandra literature, coupled to a transformer that encodes poisson-distributed photon events into a special numerical token. A metadata analyst augments this with automatically derived features describing spectral shape, time variability, noise properties, and other diagnostics. A nearest-neighbor analyst searches the aligned embedding space for similar sources with known classes and properties, providing contextual anchors for interpretation. Finally, an environment analyst uses source coordinates to incorporate information about the host galaxy or local field, exploiting the fact that location strongly constrains physical origin. Given a new Chandra observation of a source, the system autonomously computes spectral and variability summaries, identifies analogous sources in the archive, and produces expert-style natural language reports that describe likely source classes, physical parameters with uncertainties, anomalous behaviors, and recommended follow-up observations. This can be scaled to systematically scan the full Chandra archive for rare transients and strongly variable sources, showcasing how foundation models, fine-tuned on few-shot downstream tasks using numerical X-ray data and grounded in archival literature, can become practical assistants for X-ray astronomy. Our architecture also sets the foundation for scaling this AI astronomer to include other multi-wavelength data sources, creating a more robust analysis of each source.

SimAgents: Bridging Literature and the Universe Via A Multi-Agent Large Language Model System

Xiaowen Zhang
Carnegie Mellon University

As cosmological simulations and their associated software become increasingly complex, physicists face the challenge of searching through vast amounts of literature and user manuals to extract simulation parameters from dense academic papers, each using different models and formats. Translating these parameters into executable scripts remains a time-consuming and error-prone process. To improve efficiency in physics research and accelerate the cosmological simulation process, we introduce SimAgents, a multi-agent system designed to automate both parameter configuration from the literature and preliminary analysis for cosmology research. SimAgents is powered by specialized LLM agents capable of physics reasoning, simulation software validation, and tool execution. These agents collaborate through structured communication, ensuring that extracted parameters are physically meaningful, internally consistent, and software-compliant. We also construct a cosmological parameter extraction evaluation dataset by collecting over 40 simulations in published papers from Arxiv and leading journals that cover diverse simulation types. Experiments on the dataset demonstrate a strong performance of SimAgents, highlighting its effectiveness and potential to accelerate scientific research for physicists. Our demonstration video is available at: https://youtu.be/w1zLpm_CaWA. The complete system and dataset are publicly available at https://github.com/xwzhang98/SimAgents.