

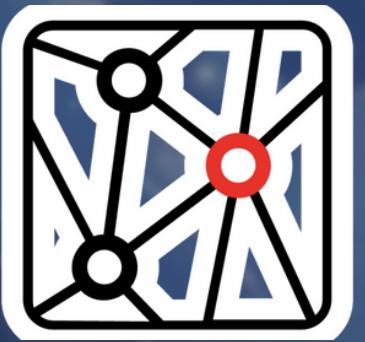
# UNLOCKING SCIENTIFIC DISCOVERIES WITH LARGE LANGUAGE MODELS IN ASTRONOMY



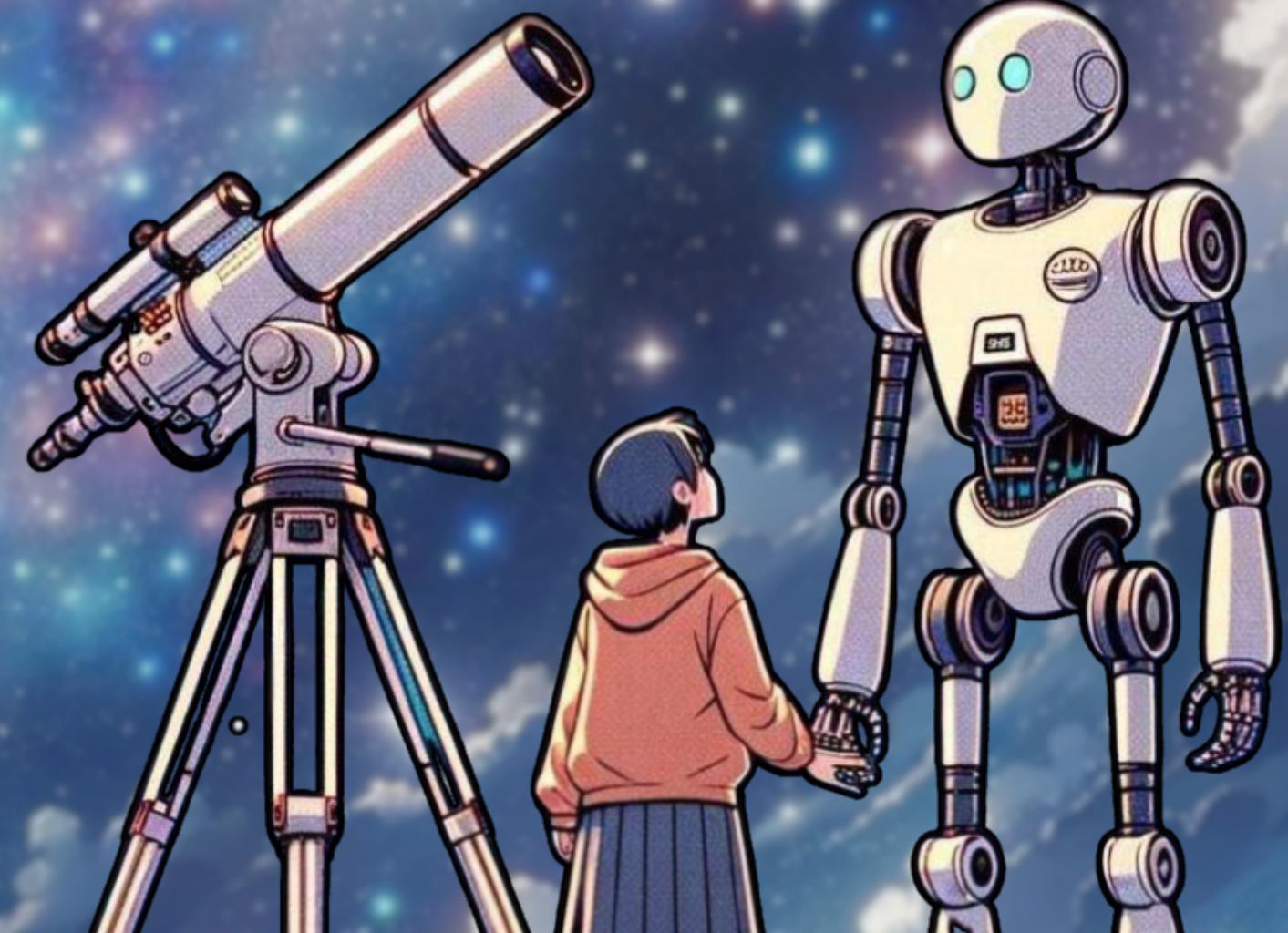
Maja  
Jablonska  
ANU



universeTBD



**MLinPL**  
CONFERENCE 2023





Alberto  
Accomazzi  
SAO/NASA ADS



David Brodrick  
ANU



Thang Bui  
ANU



Jo Ciucă  
ANU



Jesse Cranney  
ANU



Michele Ginolfi  
U.of Florence



Alyssa Goodman  
Harvard U.



Kartheik Iyer  
Columbia U.



Pranav Khetarpal  
Indian Institute of  
Technology Delhi



Sandor Kruk  
ESA Centre



Hui-ling Liu  
U.of Science and  
Technology China



Sergio José  
Rodríguez Méndez  
ANU



Jill Naiman  
U.of Illinois,  
Urbana-Champaign



Charlie O' Neill  
ANU



Josh Nguyen  
UPenn



Rui Pan  
Hong Kong U.of  
Science and  
Technology



Josh Peek  
STCI



Ernest Perkowski  
ESA



Roberta Răileanu  
FAIR(Meta)



Tomasz Różański  
ANU



Kevin  
Schawinski  
Modulos



Ze-Chang Sun  
Tsinghua U.



Yuan-Sen Ting  
ANU



Wassim  
Tenachi  
U.of Strasbourg

# **ACKNOWLEDGMENTS**

**my travel was funded by a grant from**



# **ACKNOWLEDGMENTS**

**our work has been supported by**



**CAN LLMs COME UP  
WITH UNIQUE  
SCIENTIFIC INSIGHTS  
AND HYPOTHESES?**

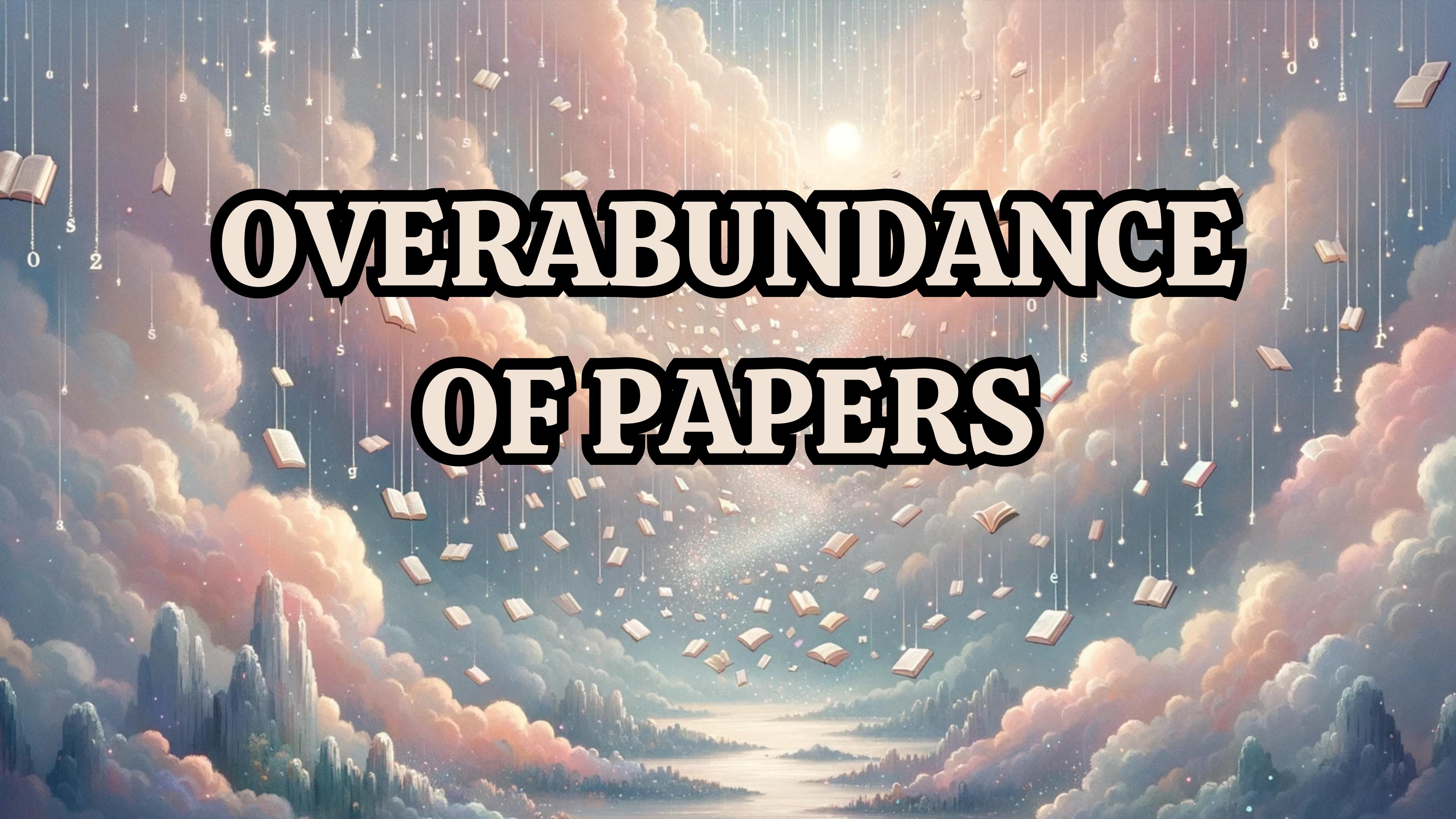
# LARGE LANGUAGE MODELS

## NO INTRODUCTION NEEDED



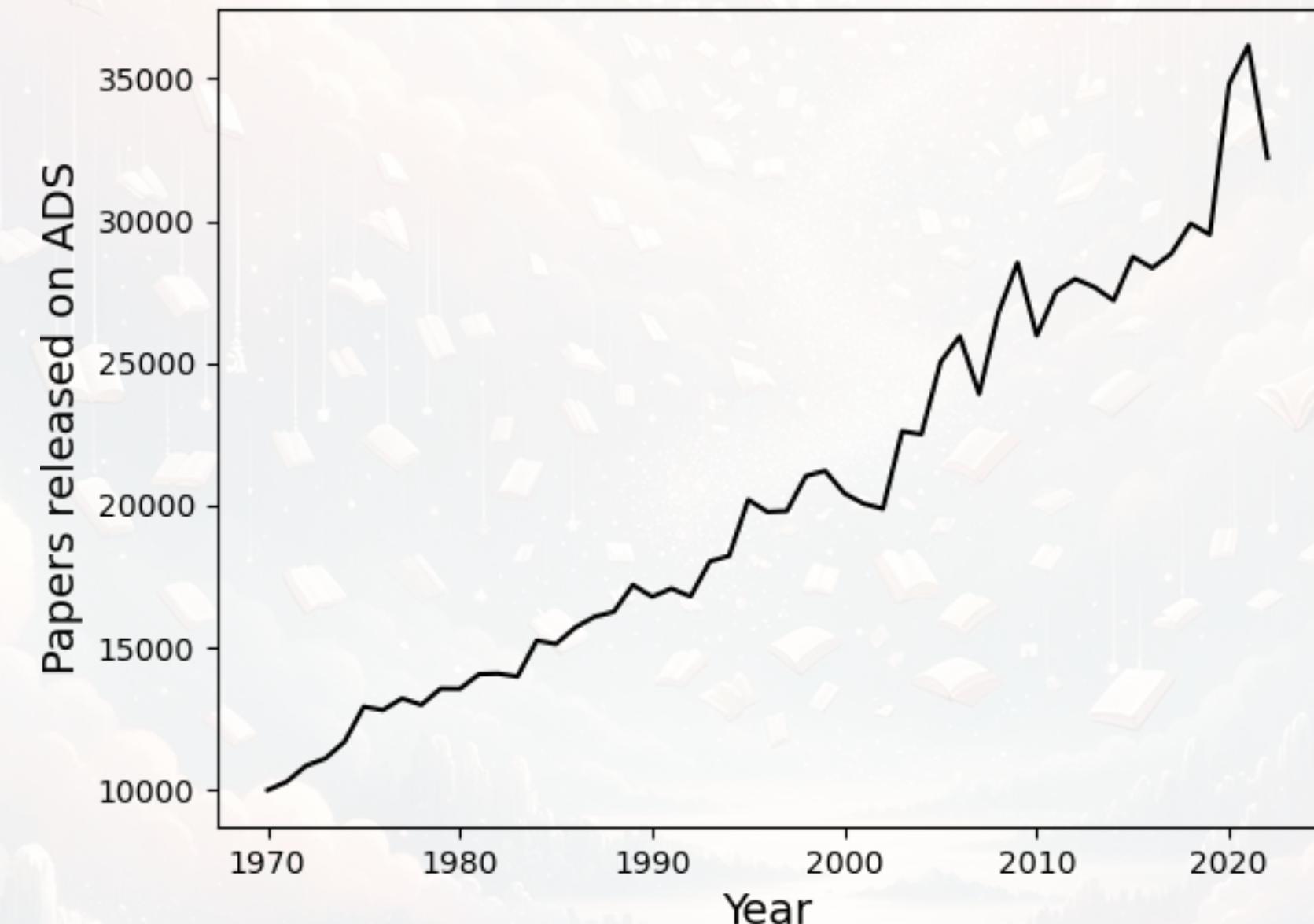
# **CHALLENGES IN MODERN ASTRONOMY**

# **OVERABUNDANCE OF PAPERS**



# OVERABUNDANCE OF PAPERS

NASA ADS: >1.3 million of refereed papers in astronomy



# OVERABUNDANCE OF PAPERS

Not necessarily contributing to faster progress?

(Slowed canonical progress in large fields of science, Chu & Evans 2021)

Inefficient metrics

Measuring similarity

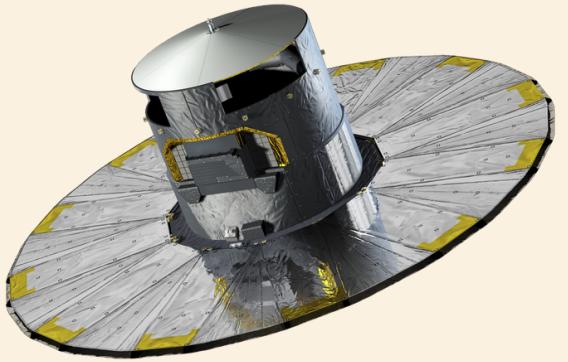
Difficult to gain an overview of the field



# UNPRECEDENTED VOLUMES OF DATA

Very often, the datasets are open source!

# **UNPRECEDENTED VOLUMES OF DATA**



**Gaia DR3 (2022)**  
**~10 TB**



**SDSS (final)**  
**~60 PB**



**SKA (per year)**  
**~300 PB**



# **SUPERCHARGING ASTRONOMERS' RESEARCH WITH LLMs**

# **CONNECTING THE DOTS**



# **DEMOCRATIZATION OF SCIENCE**



# **OUR RESULTS**



# AstroLLaMa

Nguyen et al., 2023



# AstroLLaMa

## Nguyen et al., 2023

# Text generation

**Completed by GPT-4**

The Magellanic Stream (MS) - an enormous ribbon of gas spanning  $140^{\circ}$  of the southern sky trailing the Magellanic Clouds - has been exquisitely mapped in the five decades since its discovery. However, despite concerted efforts, no stellar counterpart to the MS has been conclusively identified. This stellar stream would reveal the distance and 6D kinematics of the MS, constraining its formation and the past orbital history of the Clouds. We propose a novel approach to search for the elusive stellar counterpart of the Magellanic Stream. By employing advanced photometric and spectroscopic techniques, combined with state-of-the-art data mining and machine learning algorithms, we aim to systematically analyze and cross-reference multi-wavelength survey data to identify potential candidate stars belonging to the hypothesized stellar stream. Our methodology involves leveraging the distinctive chemical signatures and proper motions associated with stars originating from the Magellanic Clouds, allowing us to differentiate them from the foreground and background contaminants. If successful, this discovery would not only provide essential insights into the origin and evolution of the Magellanic Stream, but also enhance our understanding of the interaction between the Magellanic Clouds and the Milky Way. Furthermore, the proposed approach establishes a powerful framework for investigating other unresolved galactic structures and tracing their underlying dynamical histories.



# AstroLLaMa

## Nguyen et al., 2023

# Text generation

### Completed by LLaMA-2

The Magellanic Stream (MS) - an enormous ribbon of gas spanning 140° of the southern sky trailing the Magellanic Clouds - has been exquisitely mapped in the five decades since its discovery. However, despite concerted efforts, no stellar counterpart to the MS has been conclusively identified. This stellar stream would reveal the distance and 6D kinematics of the MS, constraining its formation and the past orbital history of the Clouds. We find a stellar stream in the MS, with a 6D velocity of 100 km s<sup>-1</sup>, and a distance of 100 kpc. The stream is 100 pc wide and 1000 pc long, and is consistent with the MS. The stream is 1000 times more massive than the Magellanic Clouds, and is likely to be a tidal stream from the Large Magellanic Cloud.



# AstroLLaMa

## Nguyen et al., 2023

# Text generation

### Completed by AstroLLaMA

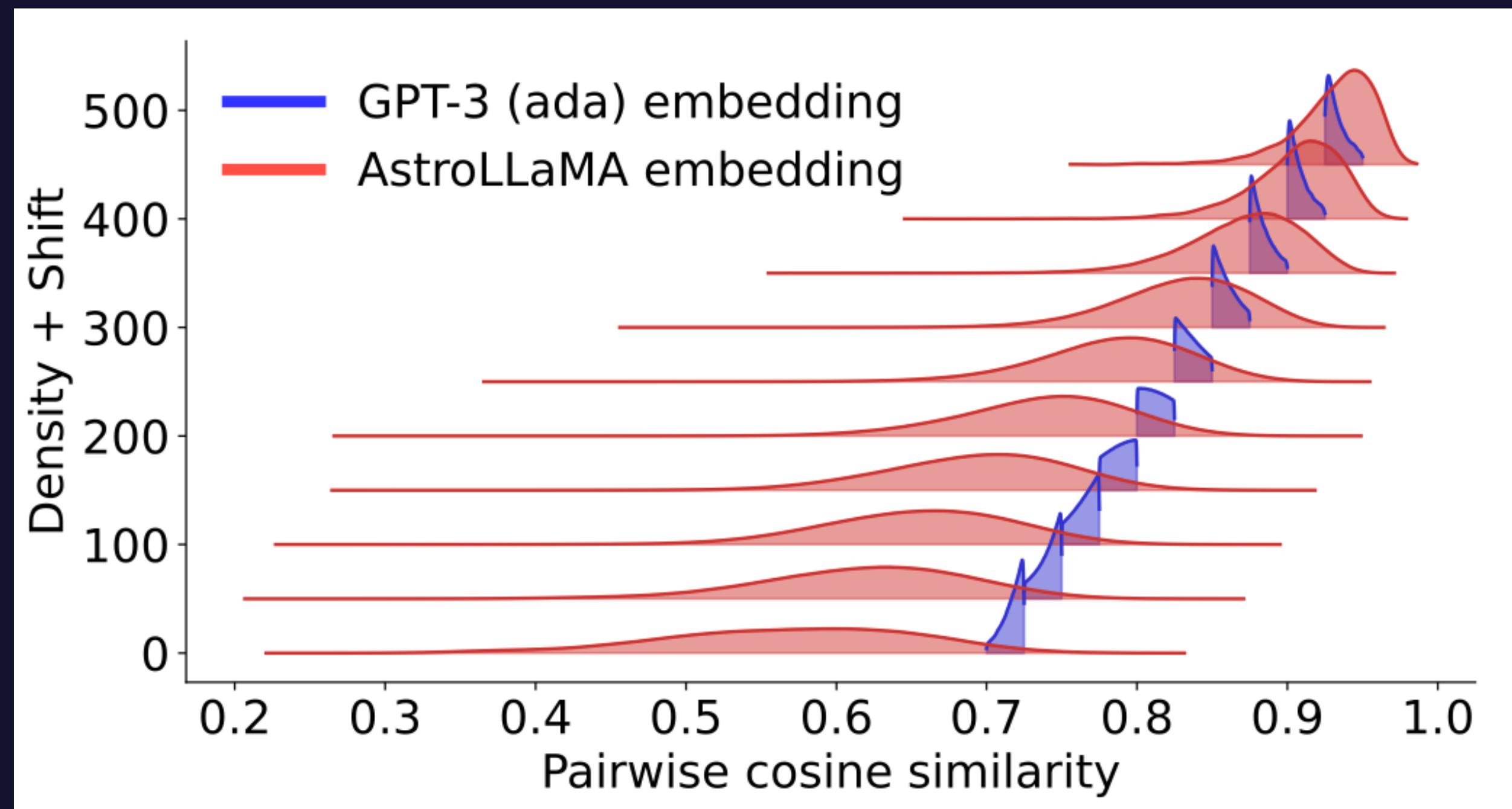
The Magellanic Stream (MS) - an enormous ribbon of gas spanning  $140^{\circ}$  of the southern sky trailing the Magellanic Clouds - has been exquisitely mapped in the five decades since its discovery. However, despite concerted efforts, no stellar counterpart to the MS has been conclusively identified. This stellar stream would reveal the distance and 6D kinematics of the MS, constraining its formation and the past orbital history of the Clouds. We 1) present a new method to select MS candidate stars from wide field optical imaging, 2) apply this method to the MS, 3) compare the results to existing catalogs, and 4) identify the most promising MS candidate stars. We apply our method to 2547 stars in the Gaia EDR3 catalog that have Gaia proper motions consistent with the MS and use these as a pilot sample to optimize our method. We then apply our method to the entire Gaia EDR3 catalog to identify 46,000 stars that are candidates for the MS, and we use the Gaia-ESO Survey to identify 13,000 of these stars that are likely to be MS candidates. We identify 266 stars with (G-K) colours and [Fe/H] that are consistent with previous estimates of the MS. These stars have proper motions and radial velocities that are consistent with the MS, and we place them on a 3D map of the MS. We find a previously unknown stellar stream, which we call the Southeastern Stream (SES), in the direction of the Clouds and MS. The SES has an M31 distance of  $32.2 \pm 0.5$  kpc and a 6D velocity of  $136 \pm 10$  km/s. The SES is about 20 times brighter in stars than the MS and about 40 times brighter in stars than the SMC. The SES is also 2.7 times more metal-rich than the MS, which suggests that it is a separate population.



# AstroLLaMa

Nguyen et al., 2023

## Embedding space





# AstroLLaMa

Nguyen et al., 2023

## Embedding space

**Paper 1:** Astrophysical gyrokinetics: kinetic and fluid turbulent cascades in magnetized weakly collisional plasma

**Paper 2:** Comment on modified Coulomb law in a strongly magnetised vaccum

*GPT-3 cosine similarity score: 78.5%*

AstroLLaMa cosine similarity score: 36.3%

**Paper 1:** A Spitzer census of the IC 348 nebula

Paper 2: Sequential and spontaneous star formation around the mid-infrared halo HII region KR 14

*GPT-3 cosine similarity score: 82.4%*

AstroLLaMa cosine similarity score: **92.8%**

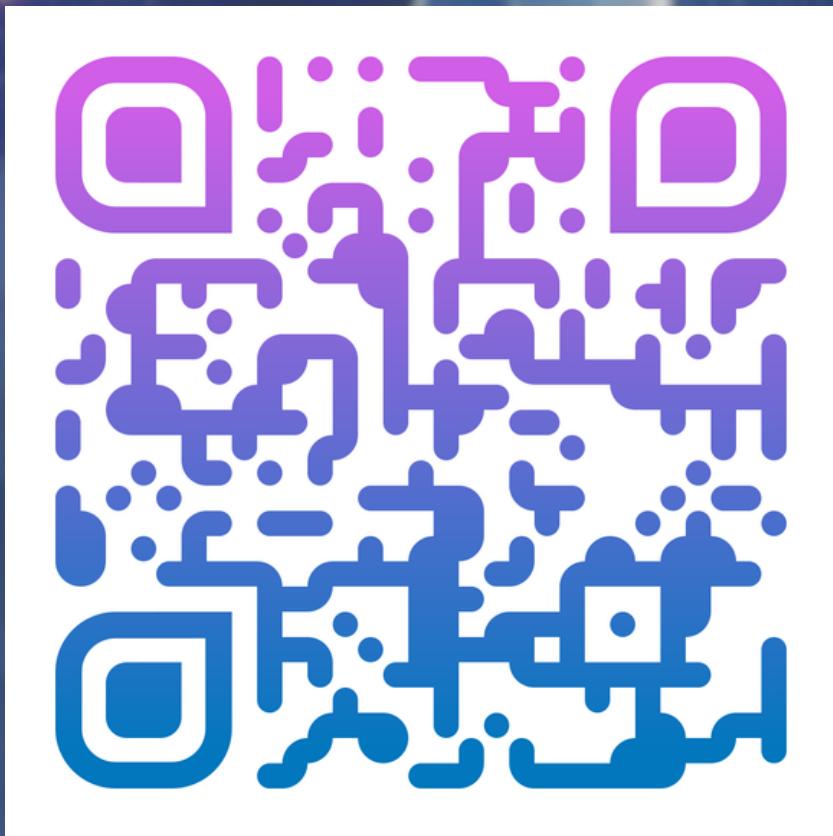
# **WHAT'S NEXT**

Training on whole papers - might help  
with knowledge gaps in selected fields

More diverse training data formats -  
Q&A, textbooks, Annual Reviews

# THANK YOU

[universetbd.org](http://universetbd.org)



arXiv

