~/ccapp-talk %

A Prompt Is Not All You Need

FROM PROMPTS TO PIPELINES WITH NEMO DATA DESIGNER

Johnny Greco | CCAPP 20th Anniversary Conference | June 2026

~/ccapp-talk %

What I've Been Up To

FROM PROMPTS TO PIPELINES WITH NEMO DATA DESIGNER

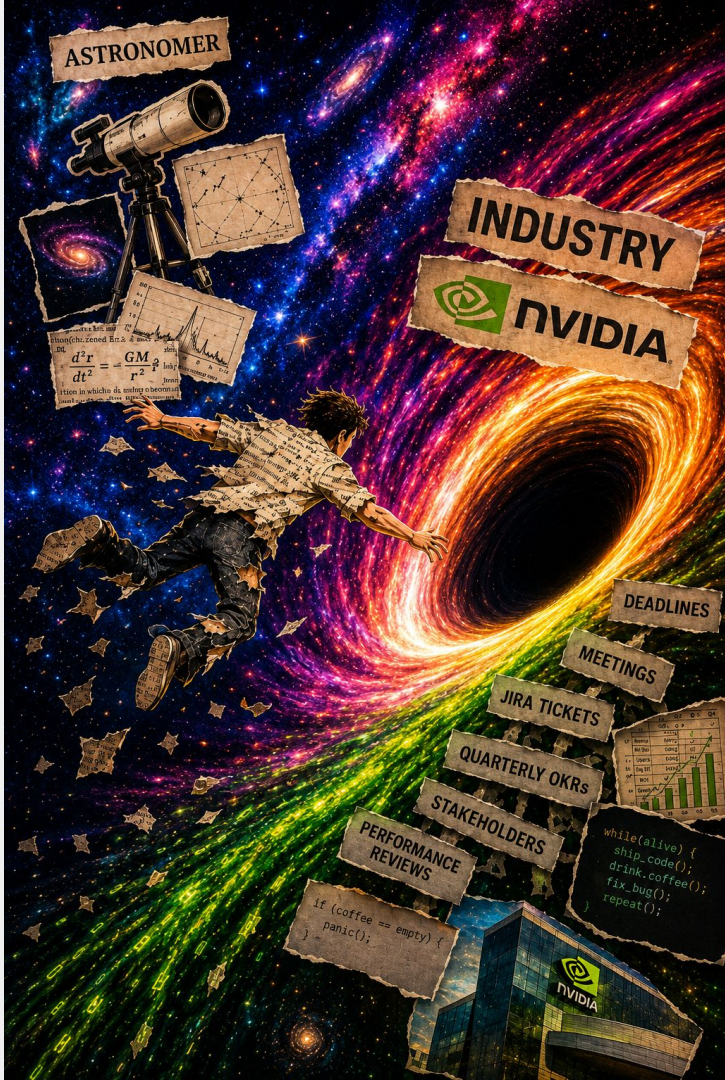
Johnny Greco | CCAPP 20th Anniversary Conference | June 2026

CCAPP, it's been awhile

Undergrad
Ohio State
2009 - 2013

Grad school
Princeton
2013 - 2018

Postdoc
CCAPP
2018 - 2021



CCAPP, IT'S BEEN AWHILE

Undergrad
Ohio State
2009 - 2013

MEASUREMENT OF THE MASS AND STELLAR POPULATION DISTRIBUTION IN M82 WITH THE LBT

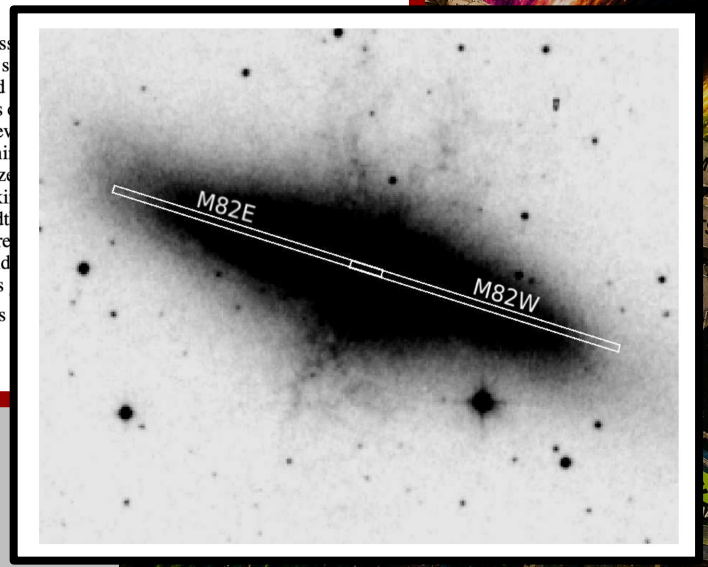
JOHNNY P. GRECO, PAUL MARTINI, AND TODD A. THOMPSON
Department of Astronomy, The Ohio State University, Columbus, OH 43210, USA; greco.40@buckeyemail.osu.edu
Received 2012 January 25; accepted 2012 July 24; published 2012 August 31

ABSTRACT

We present a *K*-band spectroscopic study of the stellar and gas kinematics, mass
of the archetypical starburst galaxy M82. Our results are based on a single s
through the *K*-band nucleus. We used the ^{12}CO stellar absorption band head
rotation curve out to nearly 4 kpc radius on both the eastern and western sides
rotation curve is flat from 1 to 4 kpc. This stands in sharp contrast to some prev
H I and CO emission-line position–velocity diagrams as evidence for a declini
the Brγ, H_2, and He I emission lines are consistent with, although characteriz
the stellar kinematics. We derived M82's mass distribution from our stellar ki
that its total dynamical mass is $\sim 10^{10}\ M_\odot$. We measured the equivalent widt
distance from the center of the galaxy to investigate the spatial extent of the re
variation in $W_{2.29}$ with radius clearly shows that RSGs dominate the light insid
likely launched from this region, where we estimate that the enclosed mass is

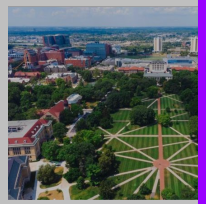
Key words: galaxies: individual (M82) – galaxies: kinematics and dynamics
galaxies

Online-only material: color figures



CCAPP, IT'S BEEN AWHI

Postdoc
CCAPP
2018 - 2021



Undergrad
Ohio State
2009 - 201

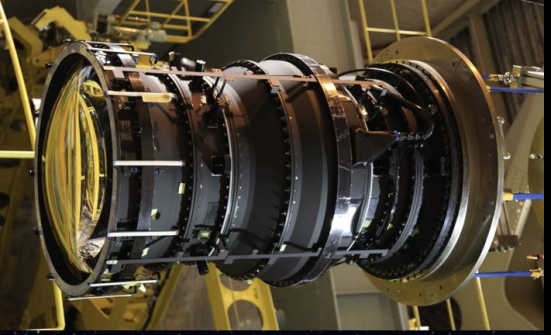
CCAPP, IT'S BEEN AWHILE

Undergrad
Ohio State
2013

Grad school
Princeton
2018

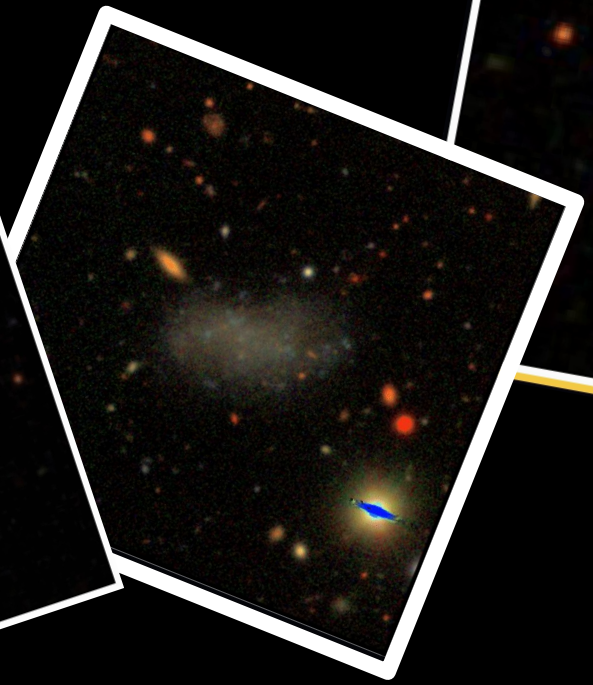
Postdoc
CCAPP
2018 - 2021

GPT 3 released

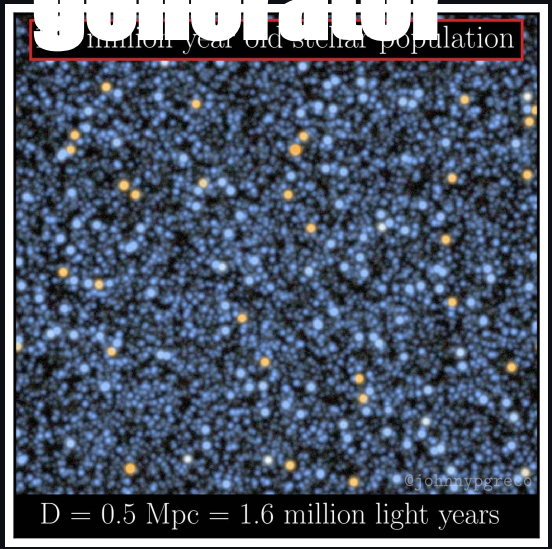


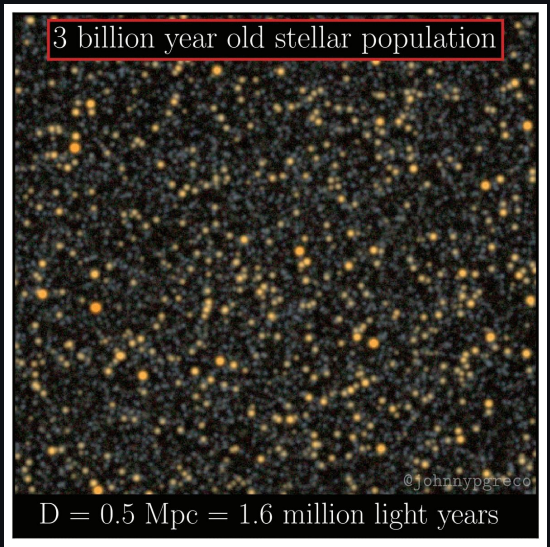
For Old Times' Sake:

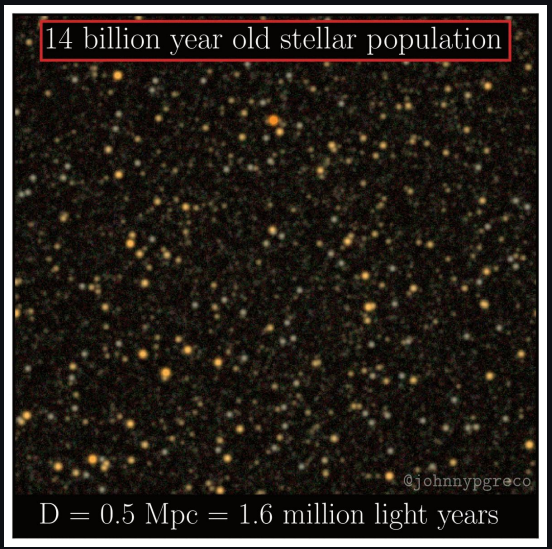
Some pretty LSB galaxy gifs



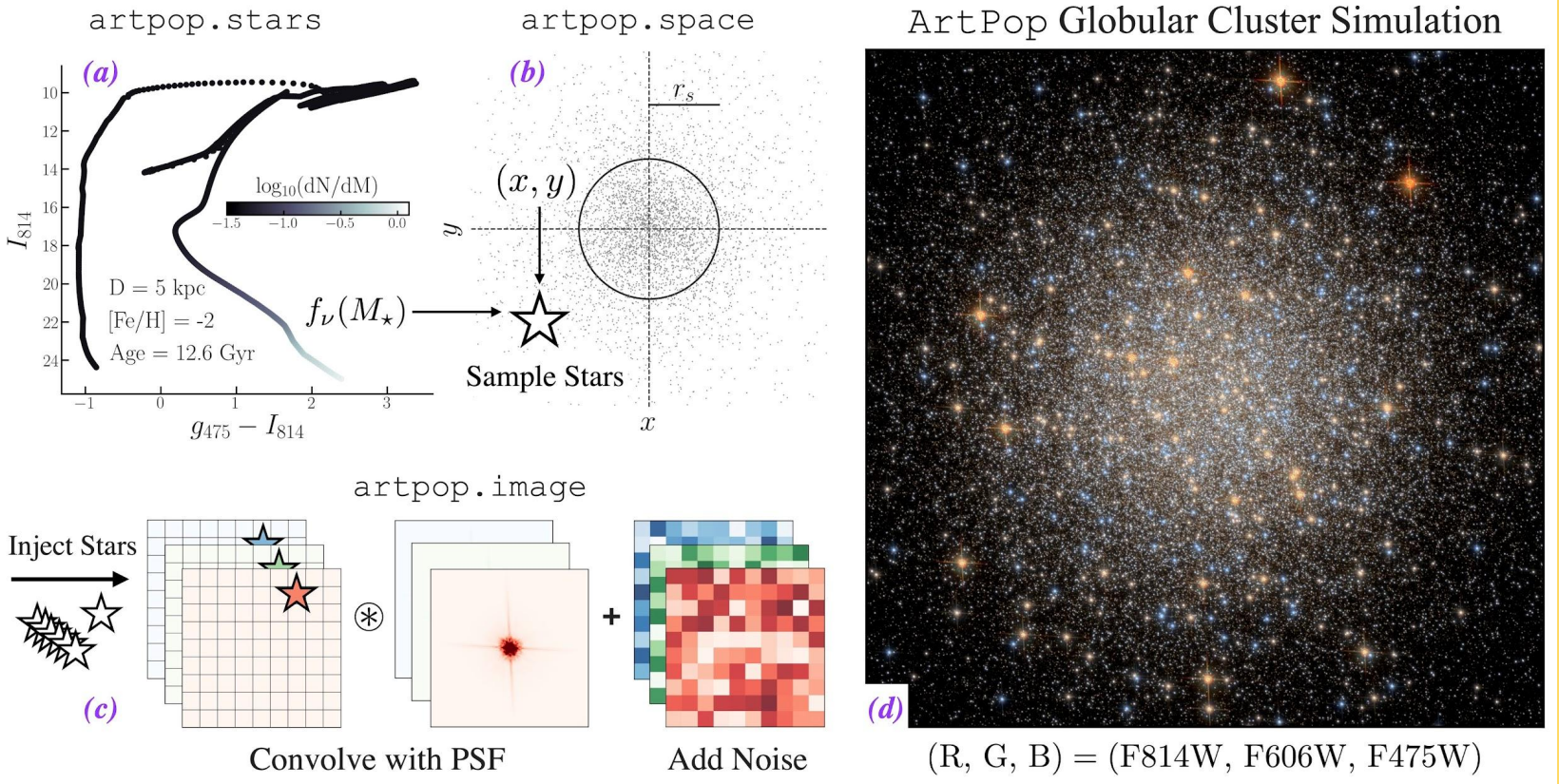
ArtPop – synthetic galaxy generator








```
$ pip install artpop
```



Greco & Danieli 2021

Synthetic Data Goes Mainstream

Synthetic Data

What is synthetic data? Why do we need it?

What is synthetic data?

Synthetic data is data generated or transformed by a process / system rather than collected directly from users in the world / the internet.

Examples :

- **Simulated scenes** for robotics, driving, or vision
- **LLM-generated text** for instruction tuning

Why synthetic data?

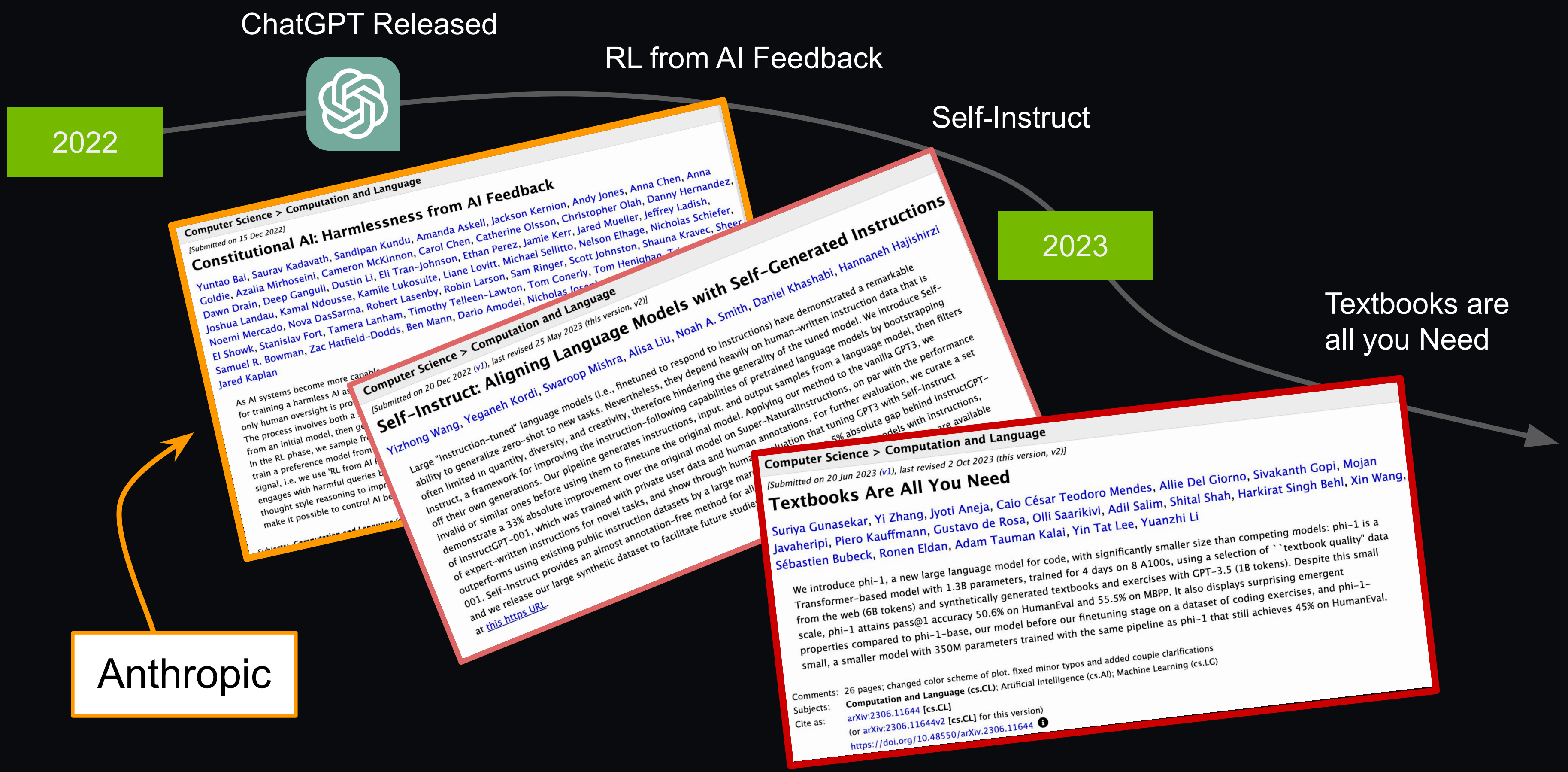
- 💸 **Cost** — expensive to collect and label
- 🎯 **Coverage** — rare and edge cases are missing
- 👀 **Bias** — real data reflects what is easy to observe
- 🔒 **Privacy + Safety** — privacy, safety, copyright
- 🎛️ **Control** — create precise data for your task

```
'''
>> Human (instruction): What is 2 + 2?
>> AI (response): 2 + 2 is 4.
'''
```

Why search (& process) for the perfect dataset when you can design it?

NVIDIA

Synthetic Data Goes Mainstream



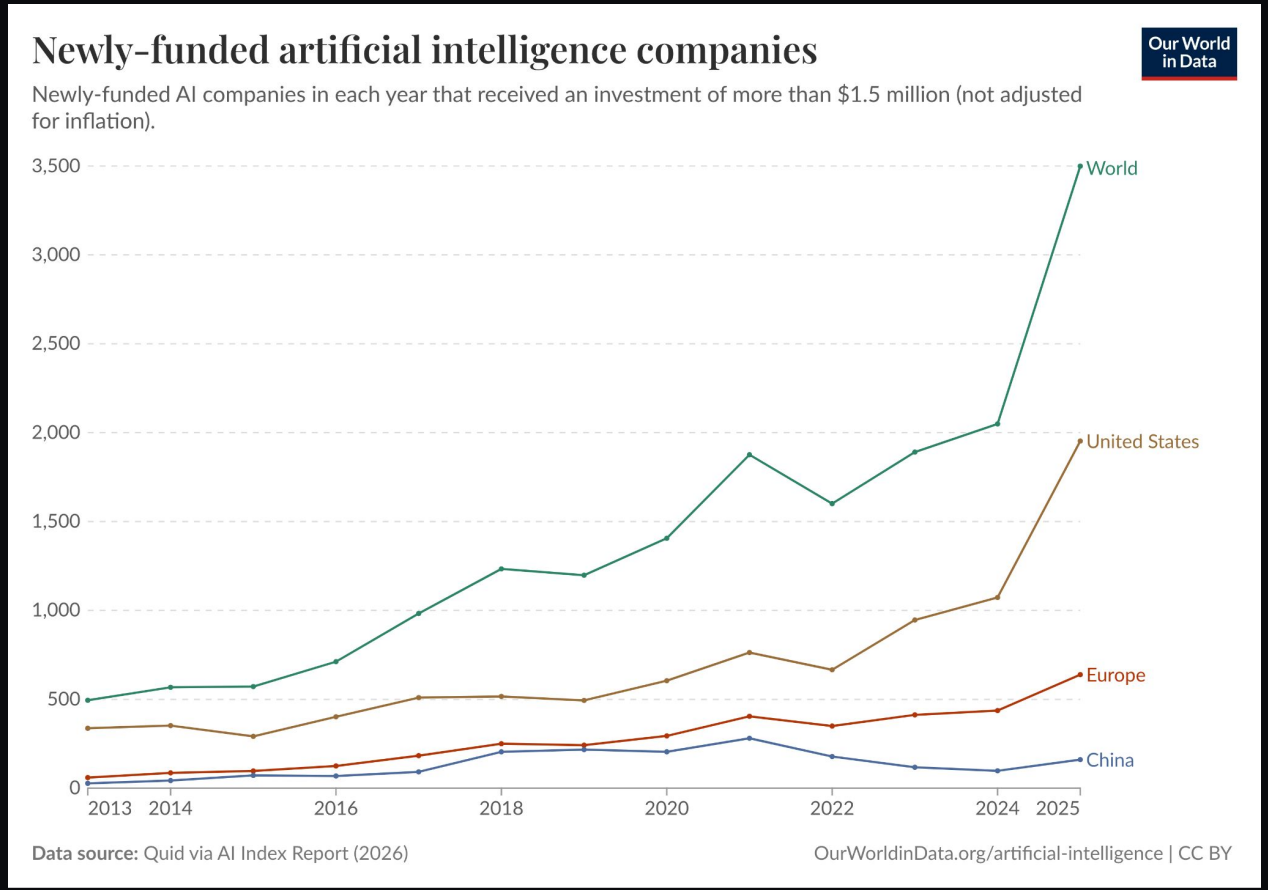
Synthetic Data Goes Mainstream

Synthetic Data Goes Mainstream

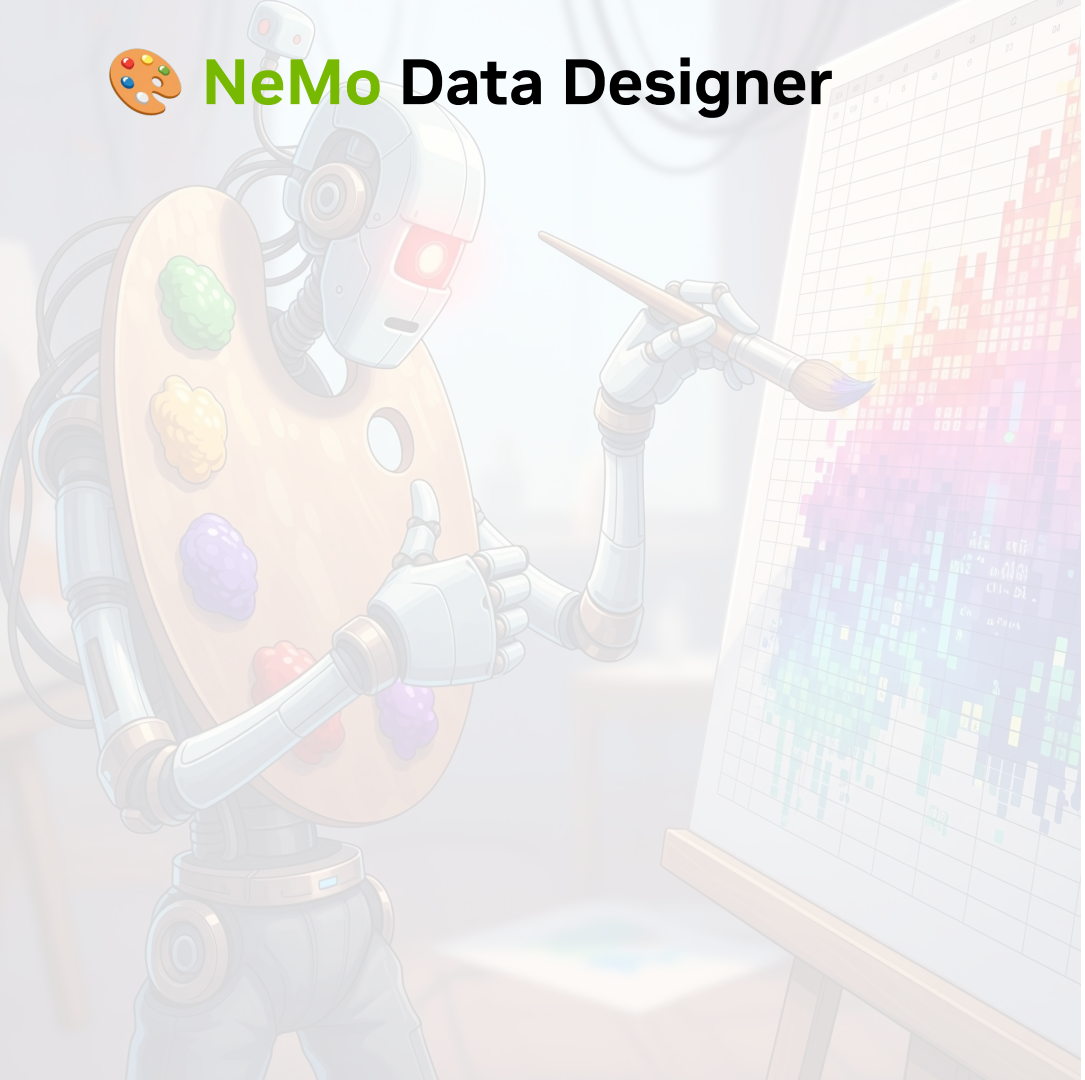
→ Applied Scientist from 2023 - 2025

Working for a Startup

Working for a Startup



🎨 NeMo Data Designer



What is it?

A **general framework** for generating high-quality ***synthetic data*** **from scratch** or based on **your own seed data**.

★ **General framework**

- Use-case agnostic
- Modular and extensible

★ **From scratch**

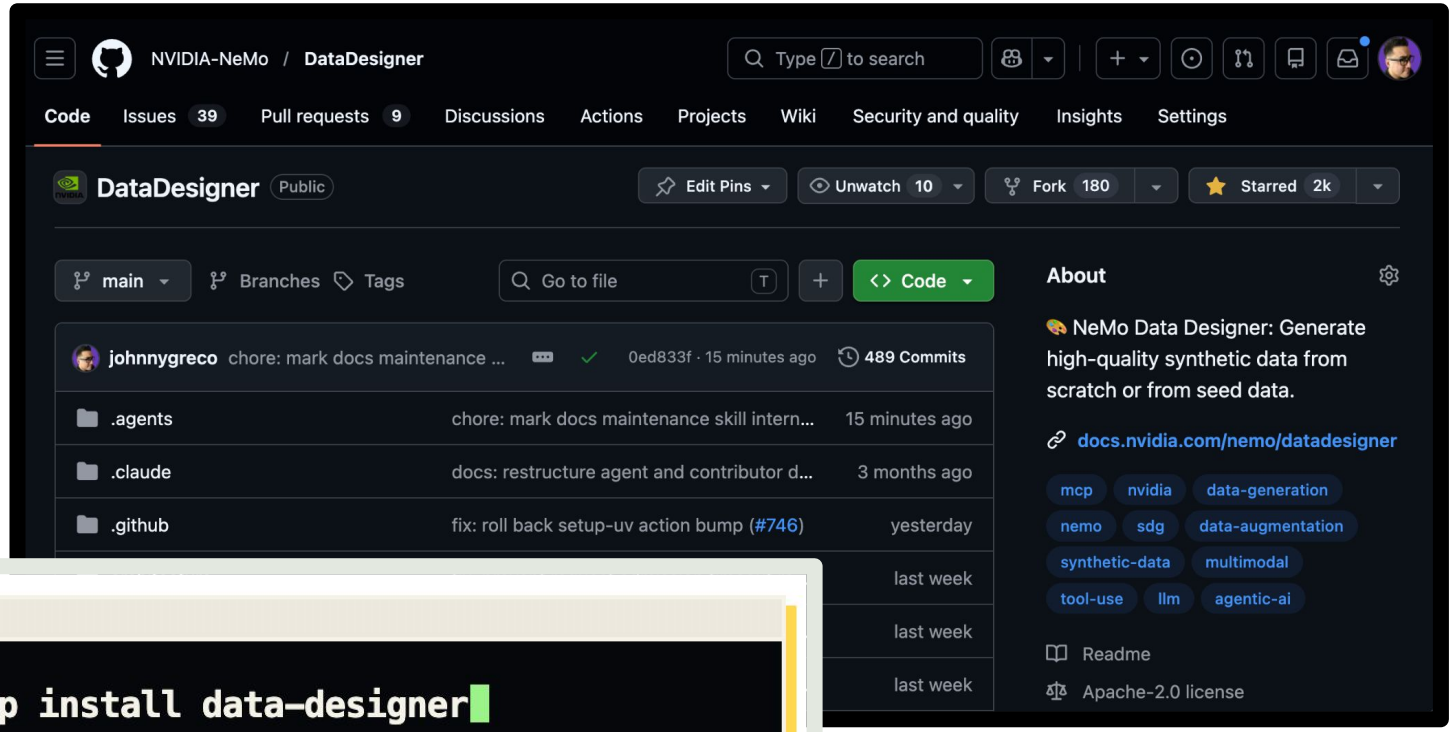
- Declarative config → *high-quality* dataset

★ **Your own seed data**

- Domain-grounded data generation

NeMo Data Designer: *What is it?*

Open Source Library: https://github.com/NVIDIA-NeMo/DataDesigner



Greco et al. in prep.

NeMo Data Designer: *What is it?*

2.5+ trillion tokens processed and *accelerating*!

NeMo Data Designer: *Why do we need it?*

```python
records = []

for _ in range(1_000_000):
    records.extend(
        [
            {"customer": llm.generate("Generate a customer profile.")},
            {"product": llm.generate("Generate a product that the customer would buy.")},
            {"review": llm.generate("Generate a review for the product.")},
        ]
    )

synthetic_dataset = pd.DataFrame.from_records(records)
```

Is this Synthetic Data Generation?

🎨 NeMo Data Designer: *Why do we need it?*

What's missing?

- Diversity, diversity, diversity!
- Correlations
- Multimodality
- Steerability
- Validation / verification
- Reproducibility
- **Diversity!**

```python
records = []

for _ in range(1_000_000):
    records.extend(
        [
            {"customer": llm.generate("Generate a customer profile.")},
            {"product": llm.generate("Generate a product that the customer would buy.")},
            {"review": llm.generate("Generate a review for the product.")},
        ]
    )

synthetic_dataset = pd.DataFrame.from_records(records)
```

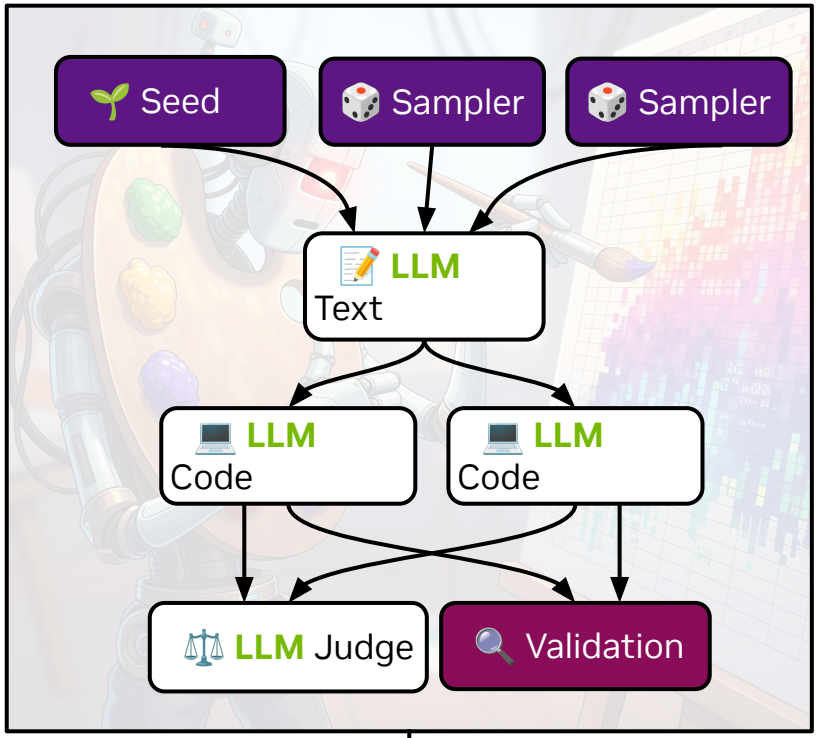
NeMo Data Designer: *How does it work?*

Column Types

- Seed
- Sampler
- LLM Text
- LLM Structured
- LLM Code
- LLM Judge
- Validation
- {{ }} Expression

Create Dataset

Dataset Builder workflow



NeMo Data Designer: Agent skill demo

Dev Notes

Dev Notes

Ask a question | Copy page | View as Markdown | More actions

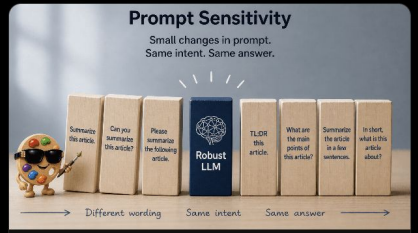
Welcome to NeMo Data Designer Dev Notes — in-depth guides, benchmark write-ups, and insights from the team building NeMo Data Designer.

JUN 1, 2026

Designing Nemotron-Personas

The compound-AI pipeline behind Nemotron-Personas seeding Nemotron training. Now open-sourced: fork it, add your data, ship personas for a...

Yev Meyer +1

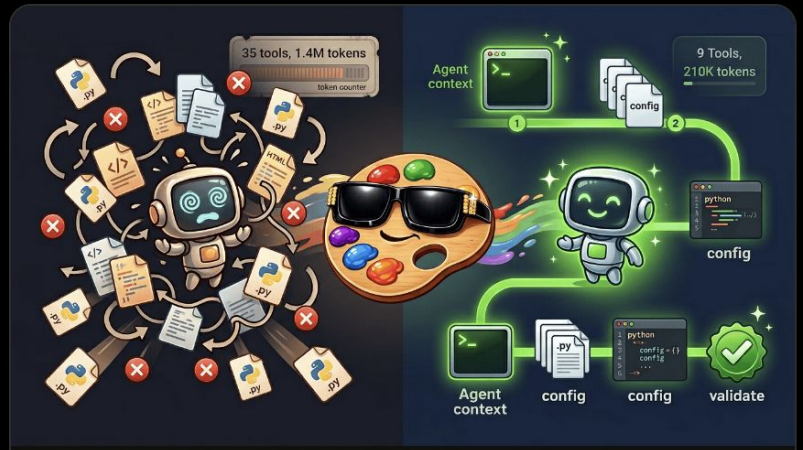


MAY 28, 2026

Mitigating Prompt Sensitivity

A Data Designer pipeline that generates thousands of regex-paired prompt preambles to manufacture format robustness — six diversity samplers, four...

Dhruv Nathawani



MAR 24, 2026

Data Designer Got Skills

A CLI and skill workflow that lets agents drive Data Designer end-to-end — leaner context, fewer tool calls, the same output.

Johnny Greco

✳ Build the project ⌥⌘1

the two accuracy=2 records.

● **Bash**(.venv/bin/python -c "
import pandas as pd, textwrap, json…)
└ ==========
[0] types of galaxies | hobbyist amateur astronomer
Q: What exactly are ultra-diffuse galaxies and why do they appear so large and faint compared to
… +74 lines (ctrl+o to expand)

The judge's reasoning is spot-on and well-calibrated — these aren't over-corrections:
- **[0]** correctly catches that the answer overstates UDG sizes as "comparable to the Milky Way" (most are dwarf-scale).
- **[1]** correctly flags genuinely wrong numbers in the answer ("tens of metres" tides, "several ms/year") and unsupported speculation.

So the harsher scores are real errors *in the answers*, surfaced by the judge — exactly the filtering signal you want for training data.

Summary of the review pass

All three issues are resolved:

Issue	Before	After
Mode collapse	Europa in 3/10, two duplicate black-hole questions	No repeated examples; questions span "what-if", observing, recent-mission (OSIRIS-REx), misconception angles
Audience mismatch	Beginner got grad-level transit Q; student got dubious infrared-bar Q	Beginners now ask simple real questions (ISS passes, "what if the Sun went supernova"); depth tracks audience
Lenient judge	29/30 scores were 4–5	Scores now span 2–5 and correctly catch real factual errors

Updated preview:

file:///Users/johnnygreco/demo-data-designer/artifacts/preview_results_20260609_232026/sample_records/sample_records_browser.html

The question_angle helper sampler does its job and is dropped from the final schema, so your output is the clean 6 columns: topic, subtopic, audience, question, answer, answer_quality.

This is in good shape. Want any further tweaks, or are you happy to finalize and scale it up?

✳ Sautéed for 3m 24s

❯ finalize it and give me the create command

▸▸ bypass permissions on (shift+tab to cycle) · ← for agents

🌯 NeMo Data Designer – Preview Record Browser

Total Records: 10, Number of Columns: 6

← Prev | Next → | Record 1

[index: 0]

Generated Columns

Name	Value
topic	Galaxies & Cosmology
subtopic	types of galaxies
audience	hobbyist amateur astronomer
question	What exactly are ultra-diffuse galaxies and why do they appear so large and faint compared to typical dwarf galaxies?
answer	**Ultra-diffuse galaxies (UDGs) are “milky-halo”–like systems that are surprisingly large—often comparable to the Milky Way in diameter—but so faint that they are hard to see even with modest telescopes.** Think of them as the cosmic equivalents of a huge, fluffy balloon that’s almost completely deflated: it stretches over many light-years, yet the amount of luminous material inside is tiny, so most of the “light” is just scattered wisps of stars spread thinly across the sky. Because their surface brightness is extremely low (often < 1 % of a typical dwarf galaxy’s), they only stand out when observed with sensitive equipment or in the data from large sky surveys. What makes them appear so big and faint? 1. **Very low star-formation efficiency.** They started out as normal dwarf galaxies, but processes such as strong tidal stripping, early bursts of star formation that blew gas out, or a lack of fresh gas have left them with only a sparse sprinkling of stars. 2. **Huge dark-matter halos.** Dynamical measurements show that many UDGs sit inside surprisingly massive dark-matter halos. The halo’s gravity can keep the galaxy stretched out over several kiloparsecs, inflating its size even though the visible stellar component is minuscule. 3. **Low surface density.** Surface brightness depends on how many stars are packed into each square arcsecond. UDGs have far fewer stars per unit area than “regular” dwarfs, so their integrated light spreads out over a large area, making each pixel look faint. In short, UDGs look large and faint because they are **low-surface-brightness, star-poor galaxies** that reside in extended dark-matter envelopes. They’re not simply “failed” massive galaxies; rather, they’re **successful** in forming dark halos but less efficient at turning that material into bright stars, giving them the “big, ghostly” appearance that challenges our usual notions of galaxy size and luminosity.

LLM-as-a-Judge: answer_quality

accuracy	clarity	audience_fit
score: 2 reasoning: The answer correctly explains that ultra-diffuse galaxies are low surface-brightness, star-poor systems residing in extended dark-matter halos and notes mechanisms such as low star-formation efficiency and tidal effects. However, it overstates their size by claiming they are often comparable in diameter to the Milky Way, which is misleading for most UDGs (they are typically dwarf-scale, only a minority reach >5 kpc). Aside from this inaccuracy, the scientific mechanisms described are largely precise and up to date.	score: 4 reasoning: The response is well-structured, uses clear headings and bullet points, and employs accessible analogies (e.g., "fluffy balloon") that help a hobbyist grasp the concepts. The explanation flows logically from definition to appearance to underlying causes, though a few phrasing choices (“milky-halo-like”) could be smoother. Overall it is easy to follow for the target audience.	score: 4 reasoning: The tone and vocabulary are appropriate for an amateur astronomer with a curiosity for deeper concepts: they avoid excessive technical jargon while still providing meaningful scientific context. The depth is balanced—neither oversimplified nor overly technical—making it a strong match for the stated audience.


~/ccapp-talk % exit
thanks, everyone

NeMo Data Designer - Moving the Needle

Data Designer was critical in developing the pre/post training datasets for Nemotron models

Goal	SDG Tokens	Results
Search Tool Calling Accuracy for Nemotron Super v3 (this workshop)	7000 records, 0.087B tokens	**0% to 31.28% on BrowseComp**
Multi-Turn Tool Call Accuracy for Nemotron Nano, Super, Ultra v3	2M Records	**2.5% to 27% on multi-turn accuracy**
Structured Output Generation for Nemotron Nano, Super, Ultra v3	0.07B Tokens	**+6.3%** on JSONSchemaBench (80.2% → 86.5%) **+8.0%** across StructEval-T (TOML, JSON, YAML, CSV and XML)
Structured Query Language Performance Nemotron Nano, Super, Ultra v3	96.5k records	**26.77% to 41.80% BIRD-bench**
Cross Domain Scientific Reasoning (InfiniByte) for Nemotron Nano, Super, Ultra v3	14.9B Tokens	Validated SDG by showing a **10% improvement** on Qwen-2.5 on **SciCode Benchmark**
STEM performance for Nemotron Nano, Super, Ultra v3 *This contributed to 2% of all pretraining data, making it the 3rd-highest-contributing SFT dataset in Super pretraining*	140B Tokens	• HumanEval Plus (**+3%** absolute) • MMLU-Pro (**+4%** absolute) • Math Reasoning (**5%** absolute) • MBPP (**+3%** absolute)
Nemotron-Personas A collection of multilingual synthetic persona datasets that support sovereign AI development across many countries and regions. Used to produce more diverse and rich data across variety of applications	Billions	• Safety data • Instruction-following data • Chat data • Multi-turn tool calling data