أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

LacteaChat: A Natural Language Interface For Exploring 2 Billion Stars from the Gaia Catalog

Mohammad Hammad, Faizan Ahmed Usman Mehmood, Abdulrahman Faisal Abdulgaffoor, Taha Noman Tasadduq, Abdul Ahad Syed Mateen, Rehan Sameer Koothalil, Reem Alghamdi

mohammad.hammad@kaust.edu.sa, faizan.ahmed@kaust.edu.sa, abdulrahman.abdulgaffoor@kaust.edu.sa, taha.tasadduq@kaust.edu.sa, abdulahad.mateen@kaust.edu.sa, koothalilsameer.firoz@kaust.edu.sa, reem.alghamdi@kaust.edu.sa

Abstract

The Gaia mission is transforming astronomy by constructing a 3D map of nearly two billion stars in the Milky Way and beyond. Despite this unprecedented dataset, the scientific value of stellar spectra remains underutilized due to the lack of large-scale interactive systems for spectral exploration. We present **LacteaChat**, a project that extends the open-source Gaia spectral visualization system with a conversational AI interface for semantic exploration of stellar spectra and astronomical structures. The system allows users to interact with the visualization pipeline using natural language commands such as navigating to regions of interest, filtering stars by physical properties, recoloring the scene according to spectral characteristics, highlighting unusual objects, and querying astrophysical populations such as white dwarfs or cool stars. Unlike conventional image-based AI systems, the proposed framework operates directly on physically meaningful spectral representations and metadata generated by the rendering pipeline, including per-pixel accumulated spectra and stellar attributes. The project investigates how large language models can serve as a coordination layer for scientific visualization by translating user intent into structured visualization, filtering, and analysis operations while relying on deterministic GPU-based backends to preserve scientific correctness. The resulting system aims to reduce the barrier to exploration of billion-scale astronomical datasets and enable more intuitive, interactive, and discovery-driven analysis workflows accessible to all.

Date: August 24, 2026

Contents

1 Introduction **2**

2 Related Work **2**

2.1 Lactea . . . 2

2.2 Conversational Agents for Data Interaction . . . 2

2.3 Retrieval-Augmented Generation . . . 3

2.4 AstroSage . . . 3

3 Methodology **3**

3.1 Exploration . . . 3

3.2 Updated Methodology . . . 4

3.3 Implementation . . . 5

4 Results **5**

5 Discussion **6**

6 Limitations and Future Work **6**

7 Conclusion **7**

References **7**

1 Introduction

The European Space Agency's Gaia mission has produced one of the largest astronomical datasets available, containing positional, physical, and spectral information for nearly two billion stars. Exploring such a large and multidimensional dataset requires visualization systems that can support efficient filtering, comparison, and analysis across many parameters. Lactea provides a web-based, multi-resolution interface for visualizing the Gaia star catalog while preserving important spectral characteristics. However, configuring the interface requires users to manually select and adjust several parameters, which can sometimes make the exploration process time-consuming and less accessible to users who are unfamiliar with the available controls.

This project introduces a natural-language interaction layer for Lactea that enables users to describe their intended visualization or analysis using ordinary language. The system interprets each prompt and converts it into a valid Lactea JSON configuration through a combination of a large language model specifically tuned for this task, predefined rules, and conventional programmatic validation. This hybrid approach allows flexible interpretation of user intent while ensuring that the generated configurations remain consistent with the capabilities and constraints of the visualization system. By reducing the need for manual configuration, the proposed system aims to make the exploration of GAIA data more intuitive, efficient, and accessible.

2 Related Work

This project builds upon previous work in astronomical data exploration, conversational interfaces, retrieval-augmented generation, and domain-specific language models.

2.1 Lactea

Alghamdi et al. [1] introduced Lactea, a web-based system for interactive exploration of the Gaia star catalogue and its stellar spectra. The system supports operations including navigation across astronomical coordinates, different sky projections, filtering, and alternative spectral colour-mapping strategies.

While Lactea provides an interactive interface for exploring the Gaia dataset, users normally interact with its functionality through explicit interface controls and parameters. LacteaChat extends this interaction model by introducing a natural-language layer that interprets a user's request and translates it into a structured Lactea configuration. This allows multiple operations, such as navigating to an astronomical region and selecting particular rendering parameters, to be expressed through a single natural-language request.

2.2 Conversational Agents for Data Interaction

Martins et al. [2] present a systematic review of conversational agents for visual analytics, examining the increasing use of natural-language interfaces for interacting with complex data and analytical systems. The review discusses approaches based on large language models, prompting, code generation, and agent-based architectures, and highlights the potential of conversational interfaces to reduce the technical knowledge required to interact with complex systems. However, it also identifies challenges including hallucination, prompt sensitivity, reliability, and inconsistent performance in domain-specific applications.

Rather than relying solely on prompting a general-purpose LLM, our approach uses an astronomy-specialized model together with task-specific training. The model is trained to translate natural-language requests into a constrained structured representation required by the application. This allows its ability to interpret domain-specific requests and follow the required

command schema to be explicitly trained and evaluated rather than relying entirely on prompt engineering.

2.3 Retrieval-Augmented Generation

Lewis et al. [3] introduced Retrieval-Augmented Generation (RAG), which combines the parametric knowledge stored within a language model with information retrieved from an external knowledge source at inference time. This approach is particularly useful for knowledge-intensive tasks because factual information does not need to be stored entirely within the parameters of the language model and can instead be retrieved when required.

RAG is particularly relevant to LacteaChat because astronomical queries can require precise object-specific information, including region names, aliases, coordinates, and object properties. Our approach retrieves astronomical information to determine the camera position and zoom directly, removing the need for AstroSage to recall these spatial parameters. AstroSage instead processes the display-related parameters, after which both components are combined to construct the final Lactea JSON configuration.

2.4 AstroSage

De Haan et al. [4] introduced AstroSage-Llama-3.1-8B, an 8-billion-parameter language model specialized for astronomy and based on Llama 3.1-8B. AstroSage was developed using continued pretraining on astronomical literature followed by supervised fine-tuning on astronomy-focused instruction and question-answer data. The resulting model substantially improves over its general-purpose Llama baseline on astronomy-specific benchmarks, demonstrating the value of domain-specific adaptation for scientific language models [4].

AstroSage provides the foundation model for LacteaChat because it already contains substantial astronomy-specific knowledge. However, AstroSage was developed primarily for general astronomical question answering and scientific reasoning rather than application-specific structured command generation or precise localization of a large collection of astronomical regions. Our work therefore further specializes AstroSage to teach the required command representation.

3 Methodology

3.1 Exploration

We are currently exploring the performance of general-purpose Large Language Models such as Llama and ChatGPT for this task. Currently, we inform the chatbot of the task via its system prompt, and it tries to estimate the parameters output as a result. However, this approach inherently has several limitations, including the fact that general-purpose LLMs do not have a high level of detailed knowledge about astronomy-related tasks. Furthermore, they have little to no accurate way of producing GAIA dataset specific information such as coordinates, spectral information etc. We have discovered that there are some LLMs which have been fine-tuned on astronomy-related datasets and perform much better.

Additionally, this simplistic approach does not change the natural underlying behavior/-tendencies of the LLM. Furthermore, the System Prompt has a limited context window, often measured in the thousands of tokens, which severely limits the amount of information that can be provided to a model at a time. Thus, we are considering training a model based on GAIA-specific data as well as examples of our intended user interactions, i.e. around 500-1000 question-answer pairs where the model learns how the user might prompt it and what kinds of

Example JSON Output from a user prompt

```
Prompt: Show me the Eagle Nebula using a temperature map
in the galactic projection

JSON Output:
{
  "camera": {
    "offset": [
      0.547094,
      0.504403
    ],
    "zoom": 0.737
  },
  "renderPipeline": {
    "projection": 2,
    "colormapStrategy": 3
  }
}
```

Figure 1 Example of a user prompt and corresponding JSON output

outputs it should generate. This form of reinforcement learning has been particularly successful in other domain-specific tasks, and we hope to see similar success.

3.2 Updated Methodology

We used a mixture of programmatic as well as AI-assisted generation to generate data using the OpenNGC dataset for the purposes of fine-tuning an LLM to perform our task. Our data is of two purposes. Firstly, we used select columns of the OpenNGC dataset to generate human-readable sentences about each object entry, with emphasis on the object's coordinates. This way we can help our target model memorize the coordinates of certain important objects, regions and galaxies. Secondly, we generated a large number of sample "Question Answer pairs" for the model to be trained upon. These consisted of a user prompt in natural language asming to visualize a specific object or region with a specific configuration, and a target JSON output that the model should ideally produce.

The JSON currently focuses on key parameters of the Lactea visualizer, including camera position (X and Y offsets), Zoom and Color Mapping.

Alongside the OpenNGC data, we scraped Wikipedia to build a catalogue of 1,258 named sky regions, covering the 88 International Astronomical Union (IAU) constellations along with other named areas. Each region has a name, an alias list, coordinates and an **angular extent**. As with the object data, we then turned this into plain sentences, one passage per region, giving each entry's name, aliases and coordinates. We intend to use these passages for Continued Pre-Training (CPT), so that the model learns the named regions a user is likely to ask about and not just individual objects.

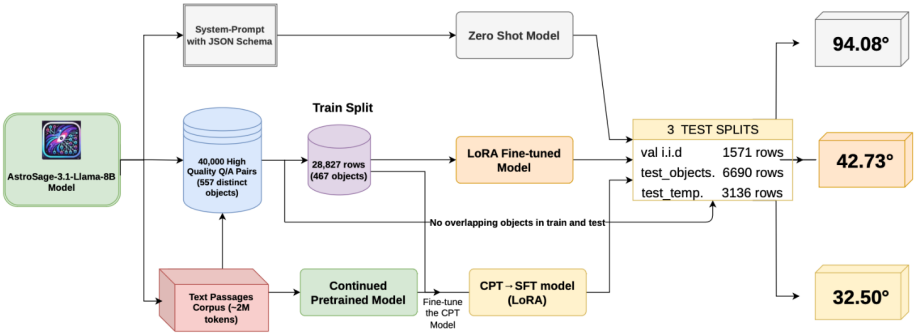


Figure 2 Pipeline comparing AstroSage zero-shot, CPT and CPT->SFT training strategies

3.3 Implementation

We used this technique to fine tune a pre-existing model known as AstroSage, which is based on the Llama 3.1 8B parameter model. AstroSage-Llama-3.1-8B is a domain-specialized natural-language AI assistant tailored for research in astronomy, astrophysics, and cosmology. It was trained on the complete collection of astronomy-related arXiv papers from 2007-2024 as well as millions of synthetically-generated question-answer pairs and other astronomical literature, making it outperform other models of comparable size in astronomy-related tasks. [4]

We want to compare two ways of fine-tuning the model. The first is CPT on the passages described above, followed by Supervised Fine-Tuning (SFT) on the question-answer pairs, so the model first picks up the coordinates and aliases and then learns the input and output format of our task. The second is SFT on its own, with no CPT. Running both lets us see whether the passages actually help, since it is possible the question-answer pairs are enough by themselves and CPT just costs training time. For both, we used parameter efficient fine-tuning rather than updating all 8 billion parameters. We used LoRA (Low Rank Adaptation) fine tuning for the model, trained on an Nvidia A100 GPU with 80 GBs of VRAM for approximately 50 minutes.

$$\theta = \cos^{-1}\left(\sin\delta_t \sin\delta_p + \cos\delta_t \cos\delta_p \cos(\alpha_t - \alpha_p)\right)$$

Figure 3 Angular separation between predicted and target positions

4 Results

Supervised fine-tuning (SFT) was performed on an NVIDIA A100 GPU with 80 GB of VRAM, with each SFT run requiring approximately 50 minutes of training. We compared two adaptation strategies: direct SFT of the AstroSage-8B model and continued pre-training followed by supervised fine-tuning (CPT→SFT). On the held-out object evaluation set, the untuned AstroSage baseline achieved a median angular error of **94.08°**. Direct SFT substantially reduced this error to **42.73 ± 4.53°** across three training seeds. The CPT→SFT approach produced a further improvement, reducing the median angular error to **32.50 ± 0.85°**, corresponding to an additional reduction of **10.23°** relative to SFT alone. These results demonstrate that supervised fine-tuning is highly effective at adapting AstroSage to the LacteaChat task, while continued pre-training provides an additional measurable improvement in the model's ability to localize previously unseen astronomical objects.

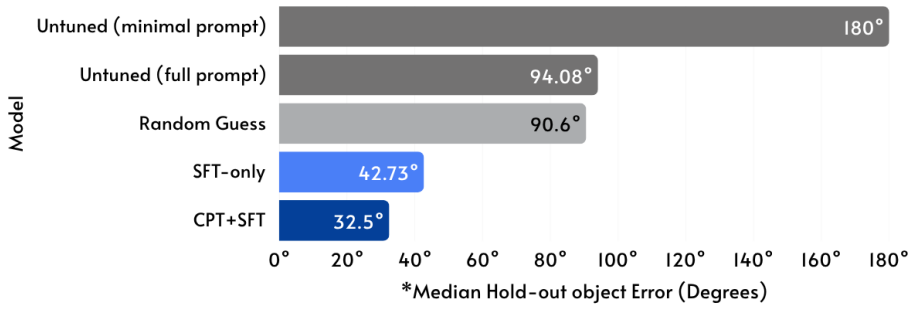


Figure 4 Median angular error on held-out astronomical objects across the untuned baselines and fine-tuned AstroSage configurations. Lower values indicate better localization performance.

5 Discussion

Although a median angular error of **32.50°** remains too large for precise astronomical navigation, it represents a substantial improvement over the untuned AstroSage baseline of **94.08°**. This reduction demonstrates that the proposed fine-tuning approaches are capable of improving the model's ability to associate astronomical objects with their corresponding locations. However, the remaining localization error indicates that further refinement is necessary before such a system can be considered reliable for scientific use. With additional improvements, a natural-language interaction layer could provide an effective and intuitive interface for navigating to and visualizing astronomical objects and regions within Lactea and similar scientific visualization systems.

6 Limitations and Future Work

Although LacteaChat demonstrates that an astronomy-specialized LLM can be adapted for natural-language interaction with astronomical data, several limitations remain regarding localization accuracy, model capacity, knowledge coverage, ambiguity handling, and experimental robustness.

Limitations of basic RAG: The current system investigates both parametric training through continued pretraining (CPT) and supervised fine-tuning (SFT), and a basic Retrieval-Augmented Generation (RAG) approach for providing astronomical information to the model. While RAG reduces reliance on the model's internal memory by supplying retrieved information at inference time, the current implementation uses a relatively simple retrieval pipeline and relies on the model to interpret the retrieved context and produce the required output. Future work should investigate **agentic RAG**, in which the LLM can actively select and use astronomical retrieval tools, query multiple catalogues or databases, resolve object identities and aliases, verify retrieved coordinates, and iteratively gather additional information when required.

Model size and computational constraints: The experiments used AstroSage-8B because training was constrained to a single NVIDIA A100 GPU. While parameter-efficient fine-tuning made experimentation with the 8B model feasible, the smaller model may limit reasoning capacity, instruction understanding, and the amount of domain-specific information that can be represented parametrically. Future experiments could investigate the larger **AstroSage-70B** model using additional GPU resources or distributed training. Comparing the 8B and 70B variants would help determine whether increased model capacity improves astronomical localization, structured command generation, and generalization to unseen requests.

Limited astronomical knowledge coverage: The astronomical-region corpus and region-of-interest dataset are constructed from a finite collection of catalogues and therefore cannot represent the full range of astronomical objects that users may query. Objects absent from the training data

are particularly challenging when their information must be recalled parametrically. Future work should incorporate larger and more diverse astronomical datasets and updated catalogues. This limitation also motivates retrieval-based approaches, since an external astronomical knowledge base can be expanded without repeating CPT or SFT whenever new objects are introduced.

Ambiguous user requests: The current system assumes that a user request contains enough information to determine the intended astronomical target and operation. In practice, objects may have multiple aliases, similar names, or ambiguous natural-language descriptions. Automatically selecting one interpretation may therefore result in a valid but scientifically incorrect command. Future systems should explicitly estimate ambiguity and request clarification when multiple plausible interpretations exist. An agentic architecture could additionally use catalogue searches and contextual information to resolve aliases before deciding whether clarification is required.

7 Conclusion

This project presented **LacteaChat**, a natural-language interaction layer designed to make exploration of the Gaia star catalog through Lactea more intuitive and accessible. The system translates user requests into structured Lactea configurations using a domain-specialized language model combined with programmatic validation. To evaluate different adaptation strategies, we compared supervised fine-tuning (SFT) of AstroSage-8B with continued pre-training followed by supervised fine-tuning (CPT→SFT). Both approaches substantially improved localization performance on held-out astronomical objects relative to the untuned model, with median angular error decreasing from **94.08°** for the baseline to **42.73 ± 4.53°** with SFT and further to **32.50 ± 0.85°** with CPT→SFT. These results demonstrate that task-specific fine-tuning can significantly improve the model's ability to interpret natural-language visualization requests, while continued pre-training provides an additional benefit for astronomical object localization. However, the remaining angular error is still too large for precise scientific navigation, indicating that further refinement is required before the system can be considered reliable for high-precision use. Overall, the results support the potential of natural-language interfaces as an effective interaction layer for large-scale astronomical visualization systems such as Lactea.

References

[1] Reem Alghamdi et al. "Lactea: Web-Based Spectrum-Preserving Multi-Resolution Visualization of the GAIA Star Catalog". In: *Computer Graphics Forum* 44.3 (2025), e70117.

[2] M. Martins et al. "Talking to Data: A Systematic Review of the Rise of Conversational Agents for Visual Analytics". In: *IEEE Access* 13 (2025), pp. 208902–208931.

[3] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 9459–9474.

[4] Tijmen de Haan et al. *AstroMLab 3: Achieving GPT-4o Level Performance in Astronomy with a Specialized 8B-Parameter Large Language Model*. 2024. arXiv: 2411.09012 [astro-ph.IM]. URL: https://arxiv.org/abs/2411.09012.