Current AI agents are limited: they are evaluated with narrow metrics, they cannot effectively handle multi-modal data and their connections, and they cannot proactively experiment and engage users. Astronomy provides an open sandbox for addressing each of these problems, and *Vela* will develop foundational advancements for the broader AI research community.

With the rapid progress in LLMs, many previously useful benchmarks have become obsolete (Bowman and Dahl, 2021), meaning that high benchmark performance does not directly translate into usability and reliability (Raji et al., 2021). Furthermore, there is limited research on how people actually interact with these models and how sub-communities benefit from them.

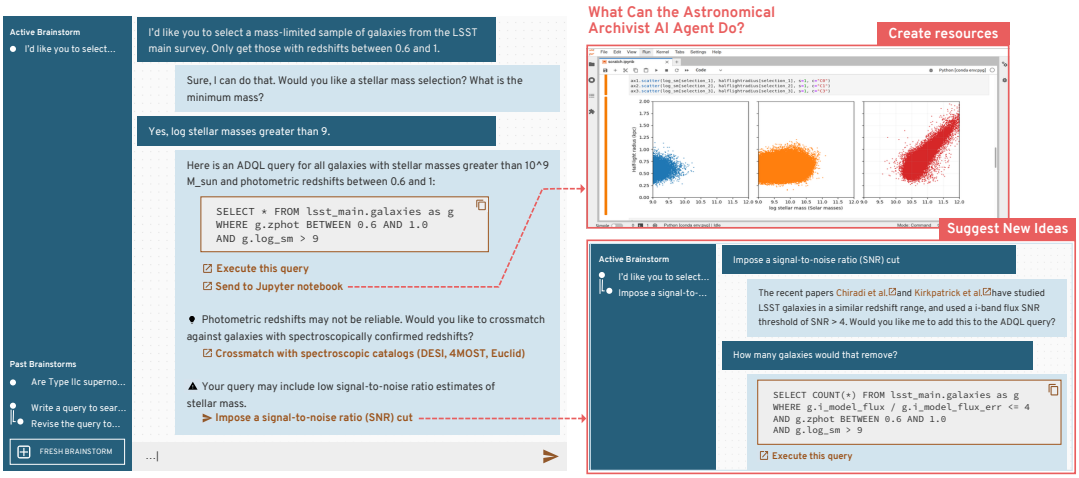

Figure 1: Example user interface for the *Astronomical Archivist AI agent* which can interact with detailed astronomical concepts and data. The AI archivist will be trained on operational metadata from STScI's systems, including logs of user-submitted SQL queries and API calls, thus learning not only the syntax of the code but also common user input errors.

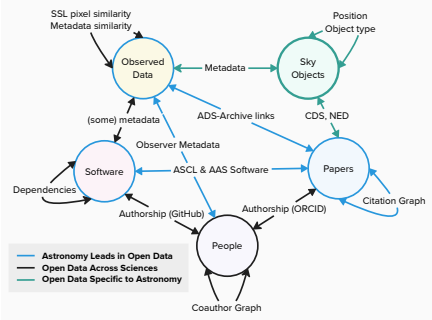

Figure 2: AI agents are powered by open data (circles), including text, software, and pixels, and the connections between them (lines). Astronomy has uniquely rich open data and documented connections (blue): not only are the observational data open, but astronomy also has a culture of sharing software in accessible repositories. The field organizes literature in an open, world-class manuscript service with papers linked to data and software.

Astronomy data is unique in that it is open and has a vibrant and active community that would partner and work with us on AI tools that we develop — becoming part of the design and development process, but more importantly, allowing for rigorous experimentation and evaluation. *Vela*, will explore a fundamental question: *How can AI transform science for the better?* *Vela* will develop an *Astronomical Archivist* LLM — which takes its name from the southern constellation named for its similarity to the sails of a ship. Like ship sails, the proposed program enables speed and exploration, opening the world of astronomical data to all.

References

S. Bowman and G. Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, 2021.

I. D. Raji, E. Denton, E. M. Bender, A. Hanna, and A. Paullada. Ai and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.