Artificial Intelligence and Large Language Models

Kelly Lockhart and the ADS Team

ADS Users Group Meeting, 16-17 Nov. 2023

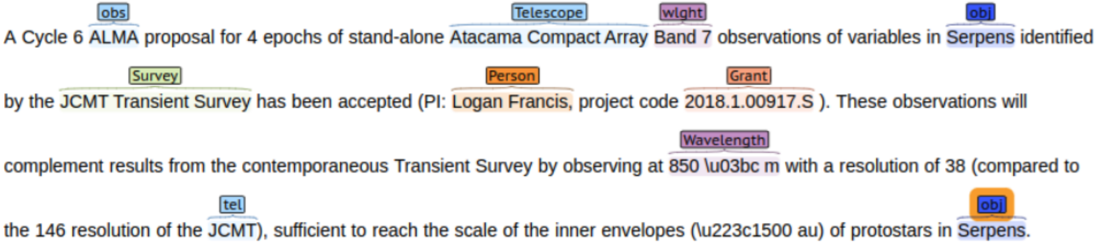
Large Language Models

- Data enrichment
    - Extension of our current machine learning efforts
- Data discovery
    - A new way of searching and synthesizing information

Data Enrichment

LLM queries are one technique of several for current back-office data enrichment tasks.

- Planetary names (in development)
- Named entity recognition (experimental)



A Cycle 6 ALMA proposal for 4 epochs of stand-alone Atacama Compact Array Band 7 observations of variables in Serpens identified by the JCMT Transient Survey has been accepted (PI: Logan Francis, project code 2018.1.00917.S). These observations will complement results from the contemporaneous Transient Survey by observing at 850 \u03bc m with a resolution of 38 (compared to the 146 resolution of the JCMT), sufficient to reach the scale of the inner envelopes (\u223c1500 au) of protostars in Serpens.

Figure: Extract of an annotated paper from the 2022 DEAL Named Entity Recognition shared task corpus.

Figure: Alkan 2023, presentation

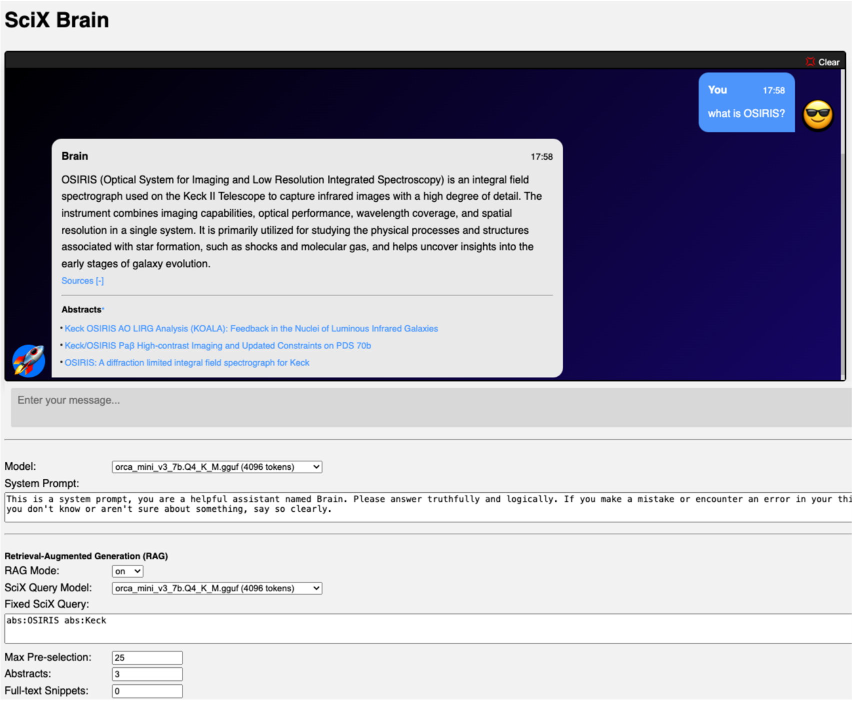
Data Enrichment: avenues to explore

- Very short (1-2 sentence) paper summaries
  - Useful for myADS email notifications
- Paper summaries written at a level for the general public, for undergrads, or for K-12 education

Figure: Alkan 2023, presentation

Data Discovery: SciX Brain chatbot

- Experimental, restricted access
    - Developed by Sergi Blanco-Cuaresma
- Test bed for LLM techniques
    - Retrieval augmented generation (RAG)
    - Comparison of various open-source LLMs
    - Architectures
    - Grammars
    - Natural language → structured Solr queries (student project)



SciX Brain

You 17:58

what is OSIRIS?

Brain 17:58

OSIRIS (Optical System for Imaging and Low Resolution Integrated Spectroscopy) is an integral field spectrograph used on the Keck II Telescope to capture infrared images with a high degree of detail. The instrument combines imaging capabilities, optical performance, wavelength coverage, and spatial resolution in a single system. It is primarily utilized for studying the physical processes and structures associated with star formation, such as shocks and molecular gas, and helps uncover insights into the early stages of galaxy evolution.

Sources [-]

Abstracts

- Keck OSIRIS AO LIRG Analysis (KOALA): Feedback in the Nuclei of Luminous Infrared Galaxies
- Keck/OSIRIS Paβ High-contrast Imaging and Updated Constraints on PDS 70b
- OSIRIS: A diffraction limited integral field spectrograph for Keck

Model: orca_mini_v3_7b.Q4_K_M.gguf (4096 tokens)

System Prompt:

```
This is a system prompt, you are a helpful assistant named Brain. Please answer truthfully and logically. If you make a mistake or encounter an error in your thi
you don't know or aren't sure about something, say so clearly.
```

Retrieval-Augmented Generation (RAG)

RAG Mode: on

SciX Query Model: orca_mini_v3_7b.Q4_K_M.gguf (4096 tokens)

Fixed SciX Query:

```
abs:OSIRIS abs:Keck
```

Max Pre-selection: 25

Abstracts: 3

Full-text Snippets: 0

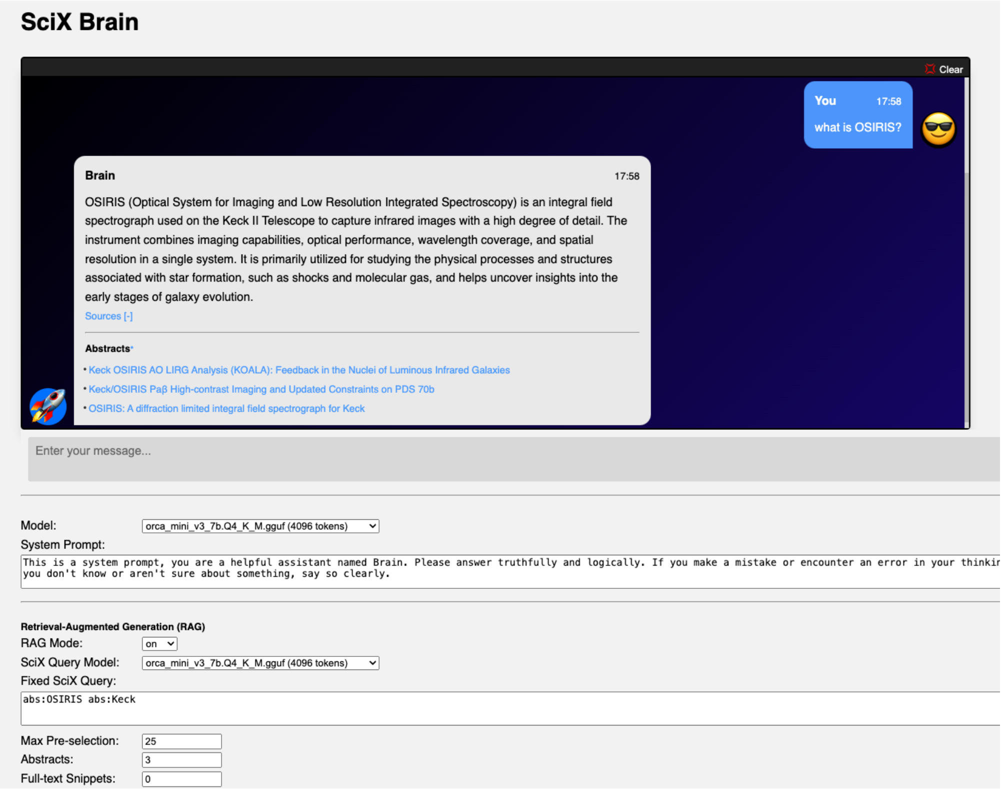
SciX Brain chatbot caveats

Concerns about opening this more widely:

- Hallucinations
  → Though some LLMs are more resilient to this
- Accuracy, reputational impact
- Exposing protected content to users
  → Though RAG mode could use only open access articles
- Cost (either purchasing GPUs or via AWS)

→ Maintaining trust, of publishers, of our users, is paramount

SciX Brain demo



SciX Brain

You 17:58

what is OSIRIS?

Brain 17:58

OSIRIS (Optical System for Imaging and Low Resolution Integrated Spectroscopy) is an integral field spectrograph used on the Keck II Telescope to capture infrared images with a high degree of detail. The instrument combines imaging capabilities, optical performance, wavelength coverage, and spatial resolution in a single system. It is primarily utilized for studying the physical processes and structures associated with star formation, such as shocks and molecular gas, and helps uncover insights into the early stages of galaxy evolution.

Sources [-]

Abstracts

- Keck OSIRIS AO LIRG Analysis (KOALA): Feedback in the Nuclei of Luminous Infrared Galaxies
- Keck/OSIRIS Paβ High-contrast Imaging and Updated Constraints on PDS 70b
- OSIRIS: A diffraction limited integral field spectrograph for Keck

Model: orca_mini_v3_7b.Q4_K_M.gguf (4096 tokens)

System Prompt:

```
This is a system prompt, you are a helpful assistant named Brain. Please answer truthfully and logically. If you make a mistake or encounter an error in your thinki
you don't know or aren't sure about something, say so clearly.
```

Retrieval-Augmented Generation (RAG)

RAG Mode: on

SciX Query Model: orca_mini_v3_7b.Q4_K_M.gguf (4096 tokens)

Fixed SciX Query:

```
abs:OSIRIS abs:Keck
```

Max Pre-selection: 25

Abstracts: 3

Full-text Snippets: 0