AstroLingua: Empowering Inclusive Astronomy through AI

A. Paulino- Afonso (IA-UPorto); Pedro Cunha (IA-UPorto); Bruno Ribeiro (Celfocus) & AstroLingua Team

Images of extragalactic objects have captured the fascination of the public, giving rise to a new generation of citizen science projects. The Galaxy Zoo project, gathering nearly one million classifications worldwide, provides a diverse dataset of galaxies in various shapes and sizes and cosmic times. The **AstroLingua** project aims at developing a comprehensive pipeline that seamlessly integrates state-of-the-art image analysis techniques and natural language processing (NLP) to deliver detailed descriptions of galaxies based on their morphology and environmental characteristics.

1. Use Galaxy Zoo [1] classifications to generate base template captions. Then use a generic LLM to create variations for data augmentation

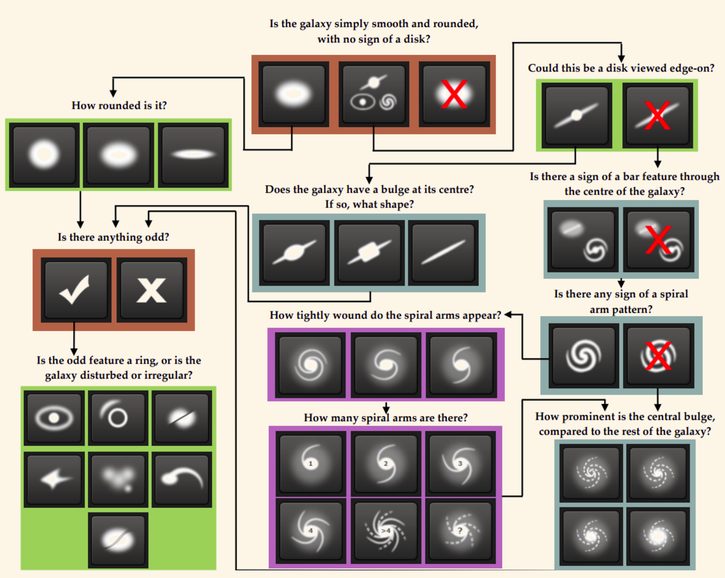


Figure from Willet et al. 2013, detailing the flowchart of classification of galaxieas

```
Uma galáxia {class} de cor {color} e que {details}
```

Llama 3.1

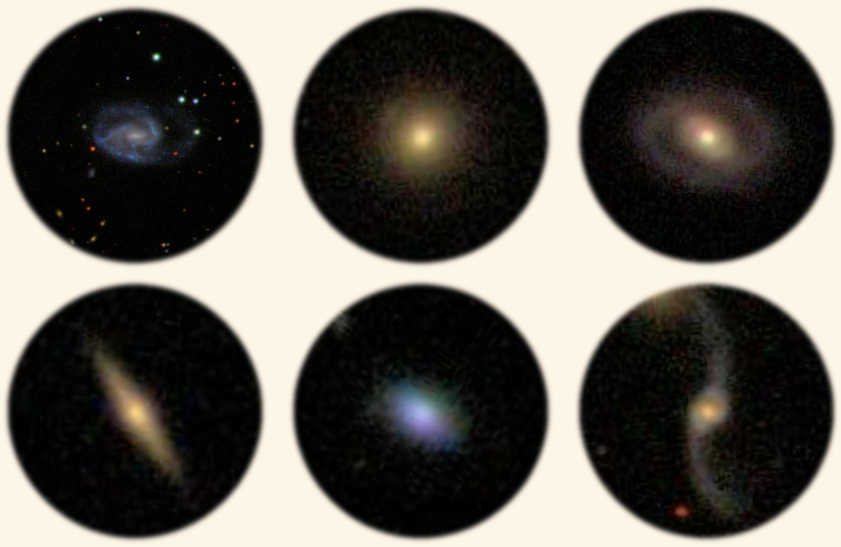
Uma galáxia de forma espiral ....

Esta é uma galáxia espiral ...

...

Estamos a ver uma espiral...

2. Collect Images from Galaxy Zoo SDSS



3. Fine-tune Image-to-text model

Image-Caption Pairs

Galáxia espiral...

Galáxia elíptica...

Galáxia irregular...

\+

Neural Net Architectures

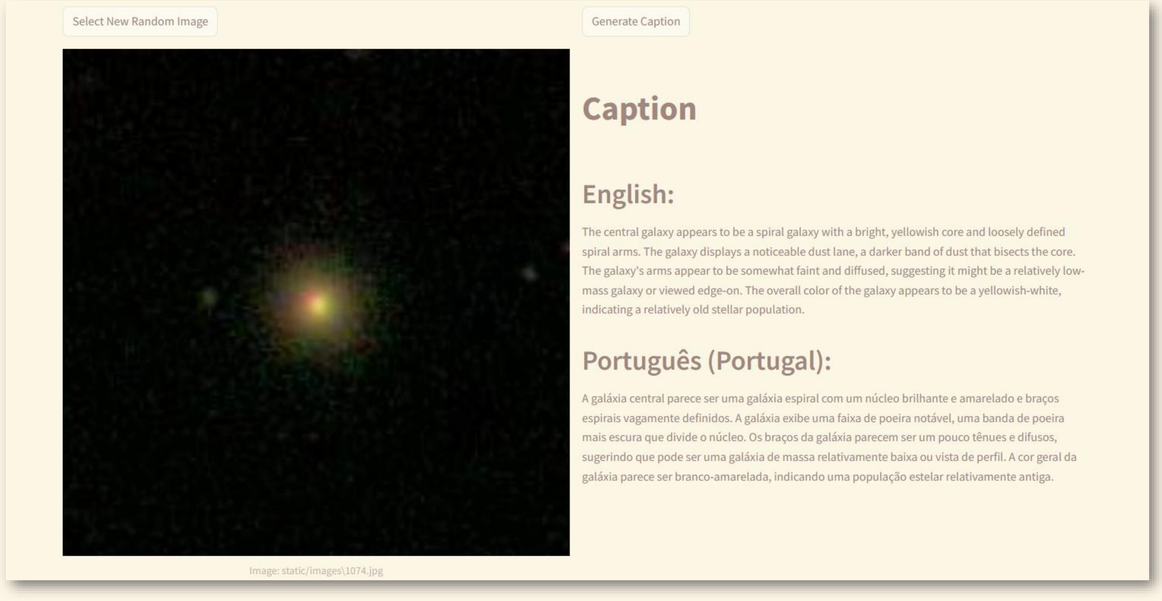
- CLIP [2]
- GIT [3]
- Qwen-VL [4]

=

ASTRO LINGUA MODEL

4. Make the model publicly available for download and through a web application for dissemination

Try it yourself here



REFERENCES: [1] Willett K. W. et al., 2013, MNRAS, 435, 2835; [2] Radford A. et al., 2021, arXiv, arXiv:2103.00020; [3] Wang J. et al., 2022, arXiv, arXiv:2205.14100; [4] Bai J. et al., 2023, arXiv, arXiv:2308.12966

ACKNOWLEDGEMENTS: A.P.A. acknowledges support from the Fundação para a Ciência e a Tecnologia (FCT) through the work Contract No. 2020.03946.CEECIND and support from the Rede Nacional de Computação Avançada (RNCA) through the project CPCA-IAC/AV/594693/2023.