Contribution ID: **125** Type: **Talk (15min + 5min)**

Generative AI for Research Data Processing: Lessons Learnt From Three Use Cases

Wednesday 26 February 2025 14:40 (20 minutes)

Generative AI has generated enormous interest since ChatGPT was launched in 2022. However, adoption of this new technology in research has been limited due to concerns about the accuracy and consistency of the outputs produced by generative AI.

In an exploratory study on the application of this new technology in research data processing, we identified tasks for which rule-based or traditional machine learning approaches were difficult to apply, and then performed these tasks using generative AI. We demonstrate the feasibility of using the generative AI model Claude 3 Opus in three research engineering projects involving complex data processing tasks:

1) Information extraction: Extraction of plant species names from historical seedlists (catalogues of seeds) published by botanical gardens.
2) Natural language understanding: Extraction of certain data points (name of drug, name of health indication, relative effectiveness, cost-effectiveness, etc.) from documents published by different Health Technology Assessment organisations in the EU.
3) Text classification: Assignment of industry codes to projects on the crowdfunding website Kickstarter.

We present the lessons learnt from this study:
1. How to assess if generative AI is a suitable tool for a particular use case, and
2. Strategies for enhancing the accuracy and consistency of the outputs produced by generative AI.

I want to participate in the youngRSE prize

yes

Primary authors: Dr OMETTO, Dawa (Utrecht University); Dr DE VOS, Martine (Utrecht University); Dr MITRA, Modhurita (Utrecht University); Dr CORTINOVIS, Nicola (Utrecht University)

Presenter: Dr MITRA, Modhurita (Utrecht University)

Session Classification: ML-assisted and more general data workflows

Track Classification: Research Software: AI and ML in a research context