Contribution ID: **30** Type: **Regular Presentation**

Metacurate-ML: Conceptual Comparison

Tuesday, 3 December 2024 11:05 (20 minutes)

Questions from the CLOSER DDI-Lifecycle repository will be used to assist in training a model that is capable of using questions and response domains from the metadata extraction workstream to create conceptually equivalent items from which data variables can be concorded. Approaches such as fine-tuned large language model (LLM)-based relevance scores model and vector retrieval-LLM reordering will be presented.
The session will present initial results in question concept tagging that feed into the conceptual comparison task, addressing challenges of long-tail distribution of the data, model memorisation and human annotation bias in the dataset. Higher-level machine learning (ML) limitations of identifying indeterminate tags and the notion of probability in model outputs will be explored.

Primary authors: DE, Suparna (University of Surrey); WANG, Zeqiang (University of Surrey)

Co-authors: Dr LI, Wing Yan (Justina) (University of Surrey); LUNGLEY, Deirdre; BRADSHAW, Paul (Scottish Centre for Social Research (ScotCen)); JOHNSON, Jon (CLOSER, UCL)

Presenter: DE, Suparna (University of Surrey)

Session Classification: Metacurate-ML