Automated, efficient extraction of structured metadata for radiology AI models and datasets

Abhinav Suri[1], Rohit Reddy[2], Ricardo A. Gonzales[3], Marcelo S. Takahashi[4], Charles E. Kahn[5], and Shahriar Faghani[6]

[1]David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[2]Department of Interventional Radiology, University of Miami, Miami, FL, USA
[3]Athinoula A. Martinos Center for Biomedical Imaging, Harvard Medical School, Charlestown, MA, USA
[4]Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[5]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[6]Department of Radiology, Artificial Intelligence Laboratory, Mayo Clinic, Rochester, MN, USA

Purpose

Artificial Intelligence (AI) research has increased rapidly over the past two decades. Thus discoverability of AI models and datasets used to train/evaluate AI is of utmost importance. "Model card" summary documents have been popularized to communicate findings; however, these model cards are unstandardized and do not capture critical information that clinicians and healthcare regulators need to know. Accordingly, the Radiology Ontology for AI Datasets, Models, and Projects (ROADMAP) has been established to describe models and datasets in radiology. The schema underlying this ontology can be used to create model and dataset cards if individuals extract structured information; however, doing so can be cumbersome, ultimately hampering widespread adoption. To address this shortcoming, we create and evaluate an automatic pipeline to extract structured information from documents describing radiological AI models and datasets.

Methods and Materials

We utilized a corpus of published papers from the *Radiology: Artificial Intelligence* journal and GPT-4o (v08/2024 with structured outputs) to extract 40 fields in ROADMAP's underlying schema. After trialing multiple approaches on a set of 5 articles and 2 RSNA grand challenge description documents, we created a two-step pipeline. First, GPT is prompted to summarize the article given descriptors of ROADMAP schema sub-fields. Next, the summary is passed into a structured output prompt using the schema to establish the format. We ran this pipeline on 54 articles and had researchers rate each subfield output as either being correct (2 pts), partially correct but incomplete (1), or incorrect (0). We report the time and cost required to run this pipeline along with an analysis of ratings for the extracted model and dataset ROADMAP schema data.

Results

Articles took 1 minute and 13 seconds to run (±15 seconds; ±1SD). The average cost per article was $0.046±0.003 (USD). 81.5% of extracted subfields were correct as is, 15.5% were correct but incomplete, and 3.0% were incorrect. The average number of points for the model and dataset fields were similar (1.76 vs 1.78, p=0.35, t-test) and there was no correlation between model/dataset scores and length of papers (in tokens, Pearson's R=-0.036, p=0.13, not different from R=0).

Conclusions

We show that our automated pipeline can extract information from radiological AI models and datasets in a manner that is reasonably cheap, efficient, and accurate. Future work includes validating this model on a larger corpus of data such as regulatory/healthcare product documents.

Clinical Relevance/Application

This pipeline can be used for enabling discoverability of AI models and datasets. Doing so can fuel research and product development in applications that can benefit patients.

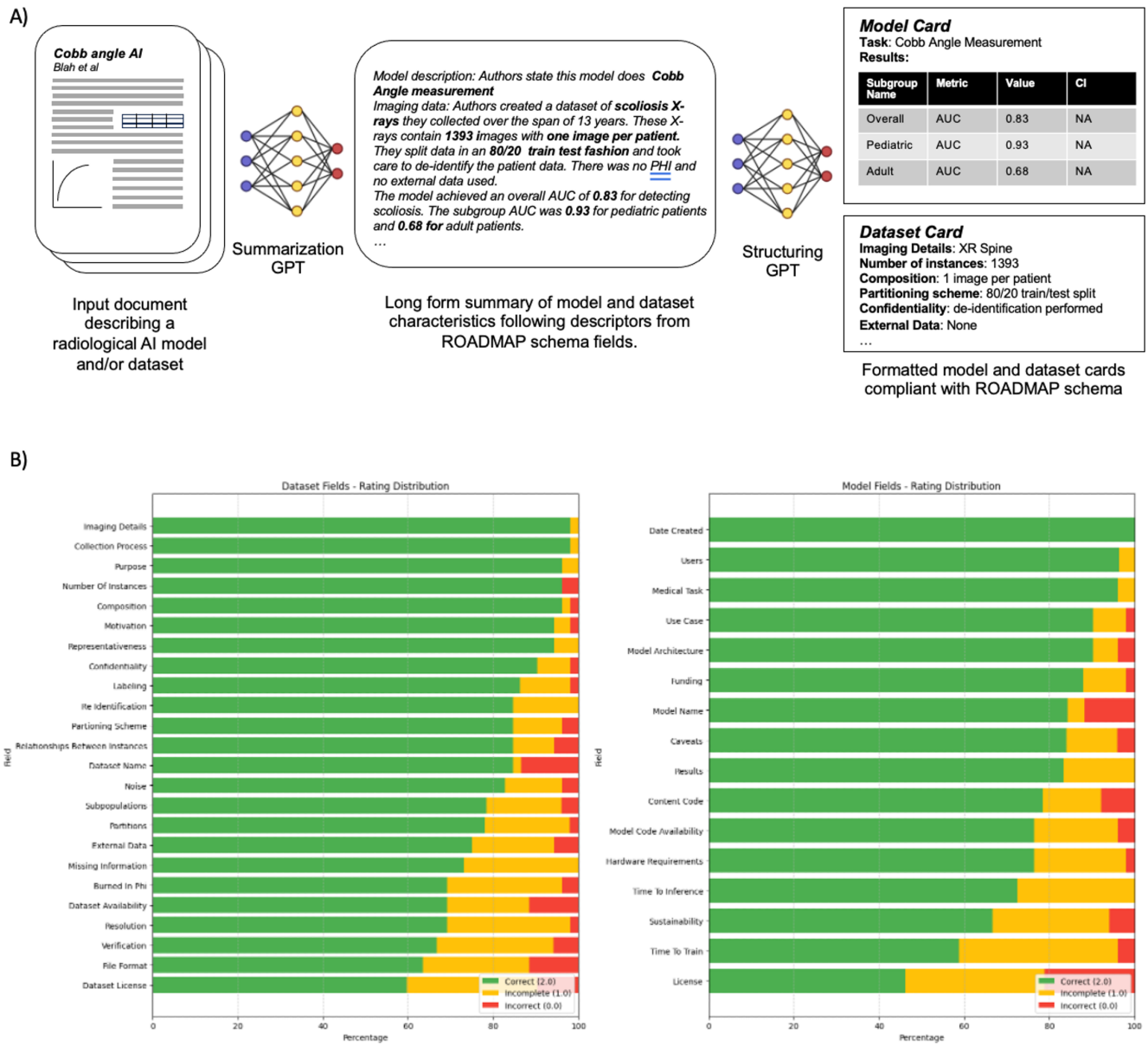

Figure 1. A) Overview of our pipeline. An input article describing an AI model and/or dataset is input to a summarization GPT which has been supplied with ROADMAP field descriptions and example field entries. The summary is then input to a structuring GPT that has the ROADMAP schema specified as its output format. The result are a model and dataset card that are compliant with the schema B) We report the distribution of rating proportions by field for the model and dataset ontologies.