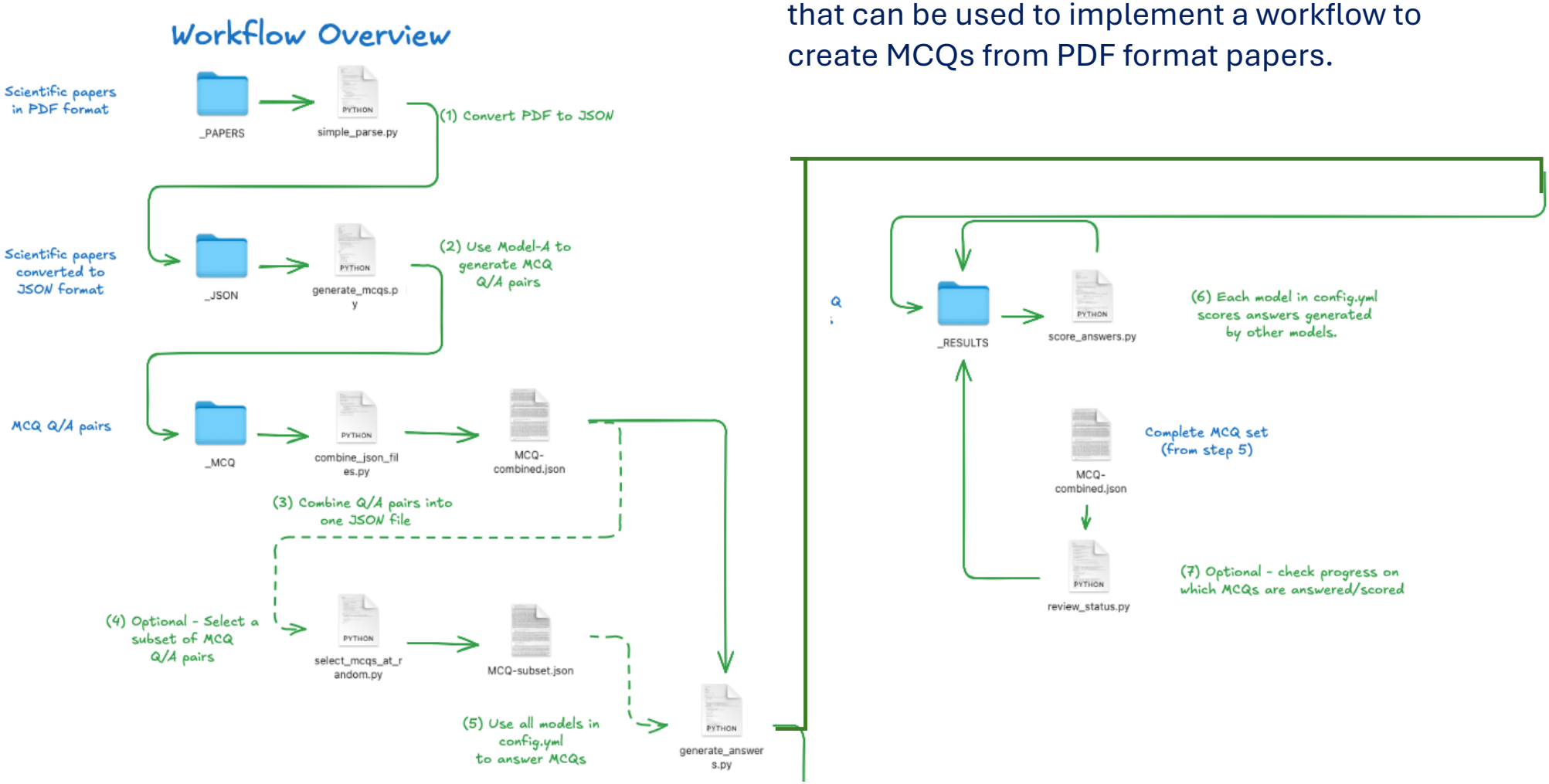
# MCQ Pipeline Update

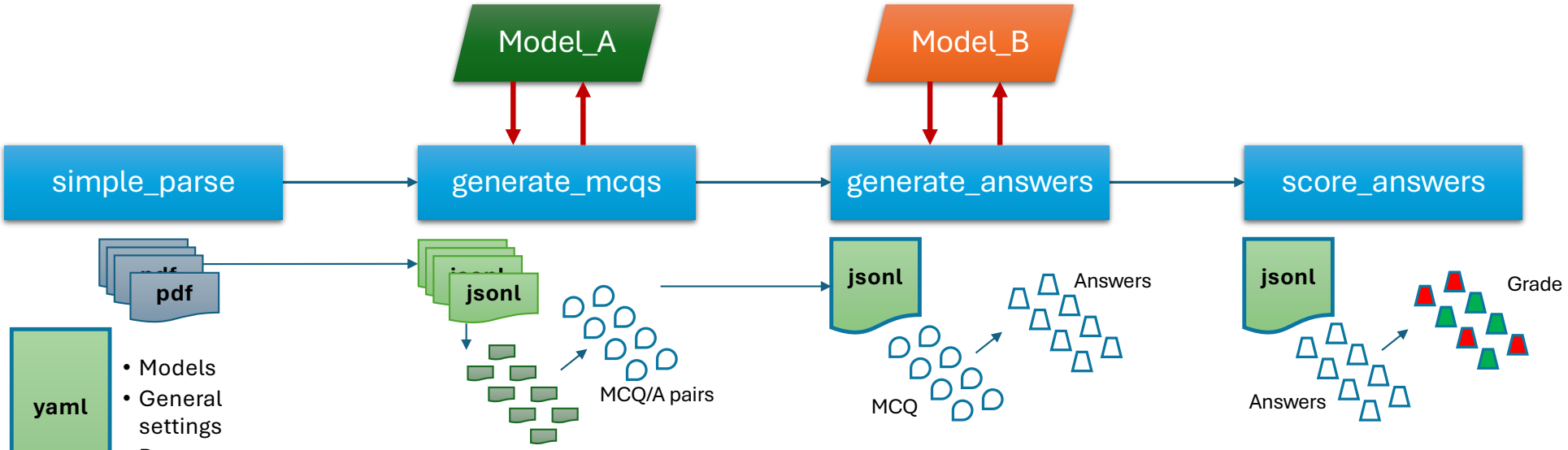
Charlie Catlett, Ian Foster, Rick Stevens

June 25, 2025

An update on a pipeline experiment to use scientific papers to create data for fine-tuning models.

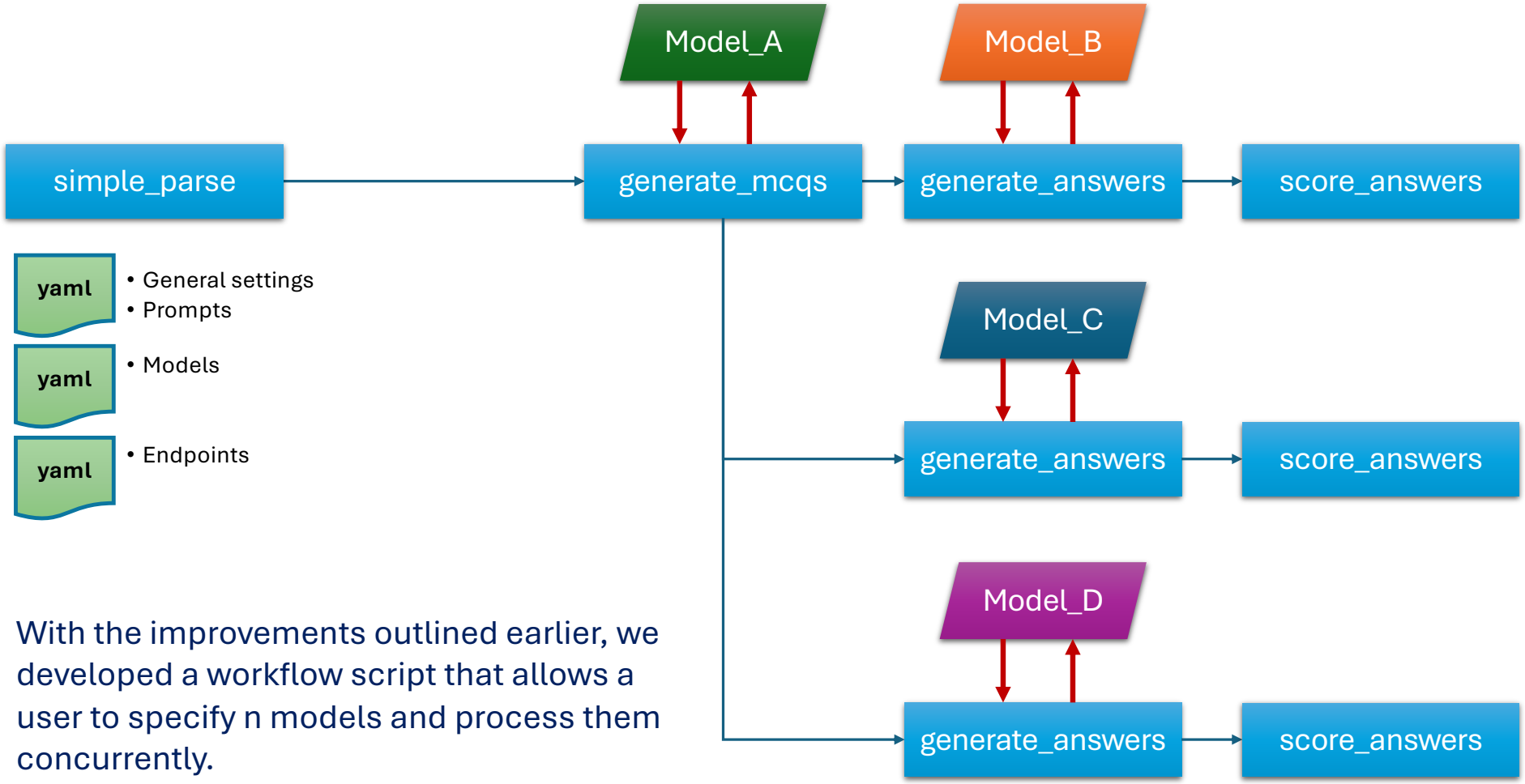
Starting point – a set of CLI tools (python scripts) that can be used to implement a workflow to create MCQs from PDF format papers.



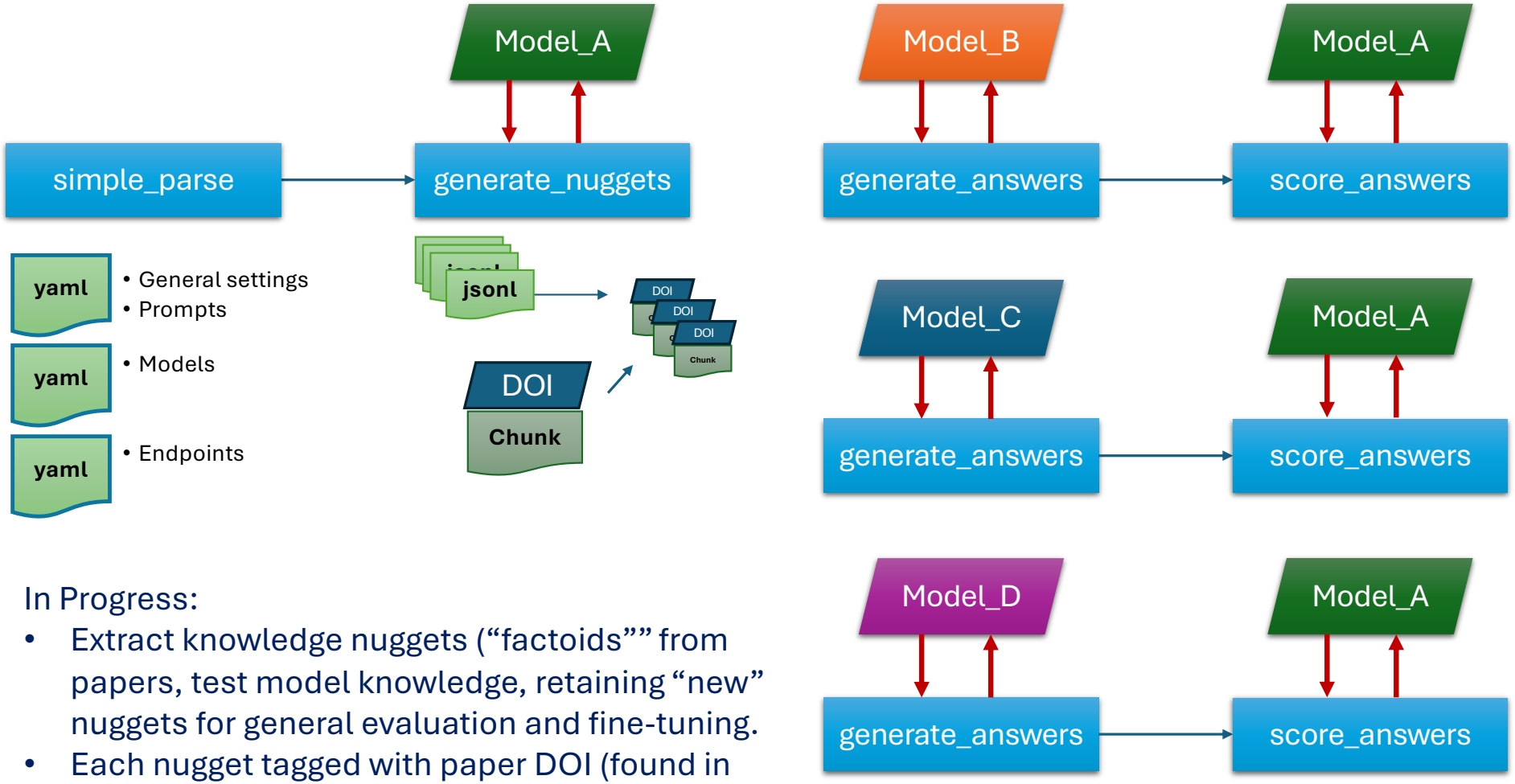



**Initial work proceeded as follows:**

1) Papers converted to JSON (text only) and broken into n-byte chunks, each of which is used to generate an MCQ. MCQ model interactions on chunks are independent, thus we parallelized this step, adding a general config setting for n-way parallel (can specify/override on command line).
2) Pull general settings, model choices into yaml config files to simplify CLI use in workflows (can be overridden by CLI options)
3) Pull prompts out of python code into yaml config file to facilitate ease of prompt engineering.
4) Pull model specifications out of python code, defining endpoints in yaml config files.
5) Redesign config handling to allow for both general (relatively static) config and local (user-specific, dynamic) config files.




With the improvements outlined earlier, we developed a workflow script that allows a user to specify n models and process them concurrently.




In Progress:

- Extract knowledge nuggets (“factoids”” from papers, test model knowledge, retaining “new” nuggets for general evaluation and fine-tuning.
- Each nugget tagged with paper DOI (found in PDF or using a model to search)