RA Task #2 (Due February 16, 2026)

Context:

This paper (Title: “How Does AI Change Drug Development? Evidence from Clinical Trial Phases and Drug Types” by Angela Kwon, Jaecheol Park, and Gene Moo Lee) examines the impact of AI in downstream drug development (clinical trials). Preliminary findings suggest that firms’ AI capability is positively associated with more refinement trials for biologics in early phases. Our current AI capability is based on AI hiring using Job Postings DB.

End Goal (in February 2026):

We will expand and triangulate the firm-level AI capability measure using **two additional evidence channels**:

1. **AI-related patents at the firm level (PatentsView)**

- Build a pipeline to identify whether biopharma firms **apply for / receive patents** that are plausibly “AI-related” (using CPC codes and/or keyword-based filtering, plus assignee mapping).
- Output: a firm-year (or firm-level) metric (e.g., AI-patent counts, AI-patent intensity, AI-patent share).

2. **Trial outcomes (ClinicalTrials.gov → Publications/Conference abstracts)**

- For clinical trials registered on **ClinicalTrials.gov**, measure whether and when trial results appear in:
    - **PubMed-indexed journal publications**, and/or
    - **Conference abstracts / proceedings** (initial focus: one major searchable conference ecosystem).
- Output: trial-level indicators (has PubMed? has conference abstract? timing lag), and a reproducible mapping method for our 9,428 NCT IDs.

Task #2:

[Merging PatentsView, DISCERN, and Clinical Trials]

A clinical trial sample is provided as a .csv file (including NCT ID, starting year of clinical trials, company name, and gvkey). Please refer to the attached file in the email.

1. **PatentsView (Patent Applications)**

   Objective: *Construct a firm-year dataset of patent applications associated with firms conducting clinical trials*

- Include both granted and pre-grant patent tables from PatentsView
- Make sure to include tables containing ‘**application-level**’ information, not only granted patent information: e.g., ‘g_application’, ‘pg_applicant_not_disambiguated’
    - ➢ Why we focus on patent applications: because applications capture firms’

earliest innovation and AI investment decisions, which are temporally aligned with clinical trial initiation and less affected by grant delays.

- Filter the data so that the application dates fall between the year 2000 – 2025
- Retain columns indicating patent type/classification (to identify AI-related patents) and abstract (potentially for text analytics)

2. **DISCERN 2 (Reference: https://zenodo.org/records/13619821)**

Objective: *Enable firm identification and gvkey matching*

- Familiarize the datasets (refer to Data dictionary pdf file)
- Identify relevant tables and variables for:
  - ➢ Mapping firm names / assignees to **gvkey**
  - ➢ Tracking firm identifiers by **year**
- Use DISCERN 2 to match **gvkey-year** information for firms appearing in the clinical trial dataset

3. **Identifying AI-Related Patent Applications**

Objective: *Identify and summarize AI-related patent activity across firms conducting clinical trials*

- Conduct EDA (exploratory data analysis) on all the patent applications filed by firms in the clinical trials sample
- At the firm-year level, identify AI-related patent applications using:
  - ➢ Patent classification codes, and/or
  - ➢ Keyword-based filtering in titles/abstracts
- Produce preliminary firm-year indicators such as:
  - ➢ Number of AI-related patent applications
  - ➢ Share of AI-related patent applications among all applications

[NCT ID – PubMed Linkage]

Objective: Identify whether registered clinical trials have associated journal publications indexed in **PubMed**.

Approach: Begin by searching PubMed using **NCT IDs** as identifiers. You may choose the specific implementation approach (e.g., automated queries via the PubMed E-utilities API, or a semi-automated workflow), as long as the method is documented and reproducible.

- Use NCT IDs as search terms: e.g., NCT01234567
- Query PubMed for publications that explicitly reference the NCT ID in any field

Data to store:

- nct_id, pmid_from_pubmed_search, publication_year, journal, any indication that the publication reference AI or AI-enabled methods (e.g., keywords in title/abstract)

Deliverable(s):

- Share the completed files as .ipynb