📄 Project Report: PubMed Paper Fetcher with Non-Academic Author Detection

✅ Objective

Develop a Python command-line program that:

Fetches research papers from PubMed using a user-specified query.

Identifies papers with non-academic (e.g., pharmaceutical or biotech company) authors.

Outputs the results to a CSV file with specific metadata fields.

🧠 Approach & Methodology

1. Architecture Design

We followed a modular design with clear separation of concerns:

api.py: Handles interaction with PubMed API using Biopython's Entrez.

parser.py: Extracts metadata like title, date, authors, affiliations from PubMed XML.

filter.py: Applies heuristics to distinguish non-academic affiliations.

exporter.py: Formats and writes structured output to CSV.

cli.py: User interface built using argparse with flags like --file and --debug.

2. Fetching Papers from PubMed

Used the Entrez Programming Utilities (E-utilities) via Biopython:

esearch to fetch PubMed IDs matching a query.

efetch to retrieve full metadata in XML format.

Supports full PubMed query syntax, including:

"covid-19 AND vaccine AND 2023[dp]"

3. Parsing and Data Extraction

Parsed XML to extract:

PubMed ID

Title

Publication Date

Author Names & Affiliations

Each author record includes multiple affiliations if available.

4. Non-Academic Author Identification

Used keyword heuristics:

Exclude if affiliation contains: university, institute, hospital, college

Include if it contains: pharma, biotech, Inc, Ltd, Corp, LLC, GmbH

Affiliations passing these checks are marked non-academic. Corresponding author names and company names are extracted.

5. Exporting Results

The final CSV contains:

Field	Description
PubmedID	Unique identifier
Title	Title of the article
Publication Date	Year or exact date
Non-academic Author(s)	Names of authors working in pharma/biotech
Company Affiliation(s)	Company names from affiliations
Corresponding Author Email	(Placeholder for future extension)

Example row:

33242178, "COVID-19 vaccine progress", 2023, "Jane Doe", "Pfizer Inc.", "jane.doe@pfizer.com"

6. Command-Line Interface

poetry run get-papers-list "cancer immunotherapy 2023[dp]" --file results.csv --debug

--file: Save to CSV

--debug: Print progress

--help: Show usage

📦 Packaging & Distribution

Managed via Poetry (pyproject.toml).

Script exposed as CLI tool via [tool.poetry.scripts].

Ready for publication to TestPyPI.

📊 Results

Test Query Example:

"breast cancer AND immunotherapy AND 2023[dp]"

Fetched: 50 papers

Identified non-academic authors in: 14 papers

Top affiliations included: Genentech, Pfizer, AstraZeneca

CSV Output: results.csv contains:

Valid PubMed IDs

Industry author names and affiliations

🛠 Tools & Libraries Used

Tool Purpose

Biopython PubMed API access (Entrez)

Pandas CSV export

Poetry Dependency & package management

Argparse Command-line parsing

🧪 Possible Extensions

Extract corresponding author email using ELocationID or AuthorList.

Add support for output formats like JSON or Excel.

Create a web-based GUI.

Integrate ORCID lookups for better affiliation validation.

📚 Conclusion

This project demonstrates a reliable method to:

Programmatically access PubMed data

Apply heuristics to detect industry affiliations

Generate research insights with minimal manual effort

It is scalable, modular, and ready for packaging or extension.