PubMed Pharma Papers Fetcher - Submission Report

Name: Ananthanarayanan S
GitHub Repository: https://github.com/Ananthu100/pubmed-pharma-papers-aganitha

Objective:
Build a command-line Python tool that uses PubMed API to fetch papers with authors affiliated with pharmaceutical or biotech companies, and outputs them as a CSV.

Approach & Methodology:

- Used ESearch and EFetch APIs from NCBI's PubMed
- Parsed XML response using `ElementTree` to extract metadata
- Identified pharma/biotech affiliations using keyword matching in author affiliations
- Extracted publication metadata including title, PubMed ID, publication date, corresponding author email

Features Implemented:

- Command-line tool with support for `--debug`, `--file`, and query argument
- CSV output with PubMedID, Title, Publication Date, Company Affiliation, Non-academic Author, and Email
- Python packaging and dependency management using Poetry
- Version-controlled using Git and hosted on GitHub

Environment Details:

- Python 3.13 environment (via Anaconda)
- Dependencies: `requests`, `pandas`
- Executed using: `poetry run python pubmedfetcher/main.py`

Results Summary:

- Successfully fetched and filtered papers based on pharma/biotech affiliation
- Sample output saved as `results.csv`

Challenges:

- Missing email data in some affiliations
- Poetry environment compatibility due to Python version constraints
- README.md recognition issue due to `.txt` extension initially

6. Future Improvements

- Improve company name extraction using Named Entity Recognition (NER).
- Improve email scraping with fallback sources.