Project Report: PubMed Paper Fetcher

**1. Project Summary**

The PubMed Paper Fetcher is a command-line Python tool designed to identify research papers from the PubMed database that include at least one author affiliated with a pharmaceutical or biotech company. This tool is useful for research analysts, life science professionals, and teams focusing on tracking industry-driven research and trends.

**2. Project Approach**

The tool uses the PubMed API to search articles based on a user-defined query (e.g., "cancer immunotherapy"). For each returned article, it examines the authors' affiliations and applies filtering rules to exclude academic institutions.

Affiliations containing keywords such as “university,” “college,” “hospital,” or “institute” are filtered out. If at least one author is identified as being from a company or a non-academic institution, the article is retained, and key details are extracted.

The final results, including author names, publication details, affiliations, and emails, are saved to a CSV file or printed to the console.

**3. Methodology**

* Built using Python 3.9 with modular, typed code for maintainability.
* Uses Poetry for environment and dependency management.
* Command-line interface (CLI) supports:
  + --file to save results
  + --debug for verbose logs
  + --help to show usage
* Leverages the Entrez ESearch and EFetch APIs from PubMed to fetch and parse XML data.
* Filters out academic affiliations using basic keyword heuristics.
* Stores structured results in a CSV format for easy reporting or downstream analysis

**4. Technologies Used**

| **Tool/Library** | **Purpose** |
| --- | --- |
| Python 3.9 | Core programming language |
| Poetry | Dependency and environment management |
| Requests | Making API calls to PubMed |
| Pandas | Data handling and CSV export |
| TQDM | Progress bar during API calls |

**5. Results**

The tool was tested with queries such as "cancer vaccine" and "AI in drug discovery". It successfully retrieved and filtered papers authored by individuals affiliated with companies like Pfizer, Roche, Moderna, and Genentech.

The output includes:

* PubMed ID
* Title
* Publication Year
* Non-academic Author Names
* Company Affiliations
* Contact Emails (when available)

**6. Conclusion**

This project delivers a lightweight and efficient way to extract industry-related research from PubMed. It automates the often tedious process of reading through affiliations manually and provides structured, exportable results in seconds.

Its modular design and use of public APIs make it easy to expand — whether by adding machine learning–based classification, handling pagination for more results, or integrating with bibliometric analysis tools.