PubMed-Pharma Research Tool – v0.1.0

PDF Report – July 2025

---

1. Objective

Developed a **CLI tool** to **fetch PubMed publications**, **filter non-academic (pharma/biotech) author affiliations**, and **export results** in CSV or JSON format.

---

2. Approach & Pipeline

1. **Keyword Matching (Layer 1)**
   - Implemented a high-precision filter using curated suffixes like `" Ltd"`, `" Inc"`, `" Pharma"`, avoiding academic overlaps.
   - Acts as an initial gate to quickly tag obvious industry affiliations.
2. **LLM-based Filtering (Layer 2)**
   - Added a fallback prompt using an LLM ("Layer 2 LLM Match") based on the commit from July 10 (GitHub, Commits).
   - Enhances accuracy for affiliation strings that don't match any keywords.
3. **Batch PubMed Integration**
   - Uses NCBI `esearch` → `efetch` to retrieve metadata in batches while respecting rate limits.
   - Handles PMIDs retrieval, XML parsing, and structured output formatting.
4. **Output Formatting**
   - Includes the following fields:
     - PubMed ID, Title, Publication Date
     - Non-academic Author(s)
     - Company Affiliation(s)
     - Corresponding Author Email
   - Export supported via CLI flags (`-f results.csv` or JSON format).

5. **Robust CLI & Logging**
    - Built using Poetry with command-line options and debug logging.
    - Supports `--file`, `--help`, `--debug`, and environment variable setup for LLM API (`set-envs.sh` or `.ps1`).

---

3. Methodology

- **Two-layer classification**:
    - High-precision keyword stage to minimize false positives.
    - LLM-based fallback for edge cases, requiring no fine-tuning.
- **Modular Design**:
    - Clearly separated pipelines under `src/pubmed_papers/pipe/`
    - CLI logic maintained in `main.py`.
- **Caching & Batching**:
    - Batches of 10 processed per request.
    - Rate-limit handling via sleep delays.
- **Flexible Output**:
    - CSV or JSON support.
    - Configurable output file using `-f`.

---

4. Results & Usage

- **v0.1.0 (Released July 11, 2025)**
    - Fully functional two-layer classification (keyword + LLM).
    - Sample usage:

```
poetry run get-papers-list "pfas exposure india" -f results.cs
```

    - Outputs processed records with industry-affiliated authors.
    - View result CSV on GitHub

- No formal accuracy benchmarks yet, but initial testing and manual reviews confirm effective filtering and classification on varied queries.

---

5. Limitations & Future Work

- **LLM Dependency**: Requires external API setup and is subject to network latency.
- **No Fine-Tuning**: Base LLM prompts may misclassify rare or ambiguous institutions.
- **Performance Metrics**: Precision/recall evaluation is pending.
- **Future Improvements**:
  - Add a semi-supervised ML classification layer.
  - Optionally fine-tune SciBERT or use a domain-specific LLM.
  - Add a UI for interactive result validation and debugging.

---

6. Conclusion

The **PubMed-Pharma Research Tool v0.1.0** is a robust and extendable CLI that:

- Efficiently retrieves and processes PubMed data.
- Accurately filters non-academic (industry) affiliations.
- Produces structured and exportable research metadata.

It provides a strong foundation for further automation, evaluation, and integration into larger research or analytics workflows.

Report based on the v0.1.0 release and current codebase.
For full source and contributions, visit the GitHub repository.

Thanks for reading 🫡 !!