**Project Summary**: ph\_resume\_ext

ph\_resume\_ext is a modular Python project designed for automated extraction, processing, and evaluation of resume data using LLMs and CrewAI. It supports DOCX and PDF formats, leverages custom tools and agents, and outputs standardized JSON for downstream analytics or HR workflows. The project includes an evaluation pipeline to compare extracted results with golden records for accuracy.

**Key Features**

- Automated extraction of structured data from resumes (DOCX, PDF)

- LLM-powered content extraction and cleaning

- Modular CrewAI agent/task/crew design

- Standardized JSON output for each resume

- Evaluation scripts to measure extraction accuracy against golden records

**Evaluation Results** (from evaluation\_results.csv)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Resume** | **First Name** | **Last Name** | **Email** | **Skills Precision** | **Skills Recall** | **Skills F1** |
| elegant-ms-word-resume-template | 1 | 1 | 1 | 1.0 | 1.0 | 1.0 |
| entry-level-data-scientist-resume-example | 1 | 1 | 1 | 1.0 | 1.0 | 1.0 |
| official-ms-word-resume-template | 1 | 1 | 1 | 1.0 | 1.0 | 1.0 |
| senior-data-scientist-resume-example | 1 | 1 | 1 | 0.5 | 0.75 | 0.6 |
| standout-ms-word-resume-template | 1 | 1 | 1 | 0.71 | 0.71 | 0.71 |

**Interpretation:**

- Most resumes achieved perfect extraction (F1 = 1.0) for names, email, and skills.

- Some resumes (e.g., senior-data-scientist, standout) had slightly lower skills extraction F1, indicating partial matches.

- Reasons for partial matches are line breaks and extra spaces only. Content is extracted complete.

**Conclusion**

The project demonstrates high accuracy in extracting key resume fields, with minor improvements possible in skills extraction for certain templates. The modular design allows for easy extension and adaptation to new formats or extraction requirements.