AI Software Intern – Internship Task Document

Internship Role Overview

As an AI Intern for 6 months (full-time), you will engage in research-driven development of Generative AI applications. The internship emphasizes both academic research and hands-on implementation, contributing to real product development, exploring research papers, and building internal tools.

Company: Wasserstoff

Role: AI Intern (Generative AI) – Full-Time, 6 Month Internship

Focus: Research-based implementation of Generative AI tools and applications (mix of research and real product development)

Required Skills: Python; Transformers and LLM architecture; LangChain framework; OpenAI API usage; system design fundamentals; basic database knowledge (SQL/NoSQL)

Internship Task: Document Research & Theme Identification Chatbot

Objective:

Create an interactive chatbot that can perform research across a large set of documents (minimum 75 documents), identify common themes (multiple themes are possible), and provide detailed, cited responses to user queries.

Task Breakdown:

1. Document Upload and Knowledge Base Creation:

- Allow users to upload 75+ documents in various formats including PDF and scanned images.

- Convert and preprocess scanned documents using OCR (Optical Character Recognition).

- Extract text content accurately, ensuring high fidelity for research purposes.
- Integrate and store uploaded documents in a database for reuse.

2. Document Management & Query Processing:

- Develop an intuitive interface where users can view all uploaded documents.
- Allow users to input queries in natural language.
- Process these queries individually against each document.
- Extract relevant responses with precise citations clearly indicating locations (page, paragraph, sentence).

3. Theme Identification & Cross-Document Synthesis:

- Analyze responses from all documents collectively.
- Identify coherent common themes across the documents (multiple themes possible).
- Produce a final synthesized answer clearly indicating all identified themes.
- Provide comprehensive citation mapping at a minimum document-level granularity to support each synthesized theme.

Additional Functionalities (Extra Credit):

- Enhanced granularity of citations (paragraph or sentence level).
- Visual representation or mapping interface linking citations to documents.
- Advanced filtering options (date, author, document type, relevance scores).
- Enable selection/deselection of specific documents for targeted querying.

Technical Requirements:

- Modern AI language models (OpenAI GPT, Gemini, Groq) .
    - ❖ GPT and Gemini give free credits for testing and project purposes
    - ❖ Groq is a free LLM service hosting LLAMA.
- Vector databases (Qdrant, ChromaDB, FAISS) for efficient semantic search.
- OCR libraries (Tesseract, PaddleOCR) for scanned documents.
- Python frameworks such as FastAPI or Flask for backend.
- Deployment on free hosting services if necessary (see deployment links below).

Deliverables:

- Fully functional web-based chatbot with documented code.
- Brief report explaining methodologies and technologies used.
- Demonstration video or presentation showcasing functionality and results.

Evaluation Criteria:

- **Functionality:** Objectives achievement of document research, theme identification, citation.
- **Code Quality & Structure:** Modular design, readability, proper naming, comments, version control usage.
- **Error Handling:** Robust handling of exceptions or failures.
- **Documentation Clarity:** Comprehensive README and demonstration video.

- **System Design Insight:** Scalability and deployment considerations.
- **User Interface:** Simplicity and clarity.

Presentation of Results:

- Individual document responses in tabular format. Example:

Document ID	Extracted Answer	Citation
DOC001	The order states that the fine was imposed under section 15 of the SEBI Act.	Page 4, Para 2
DOC002	Tribunal observed delay in disclosure violated Clause 49 of LODR.	Page 2, Para 1

- Final synthesized response in chat format with clear citations marked by document IDs. Example:

Theme 1 – Regulatory Non-Compliance:
Documents (DOC001, DOC002) highlight regulatory non-compliance with SEBI Act and LODR.

Theme 2 – Penalty Justification:
DOC001 explicitly justifies penalties under statutory frameworks.

Recommended Folder Structure:

```
chatbot_theme_identifier/
├── backend/
│   ├── app/
│   │   ├── api/
│   │   ├── core/
│   │   ├── models/
│   │   ├── services/
```

```
│   │   ├── main.py
│   │   └── config.py
│   ├── data/
│   ├── Dockerfile
│   └── requirements.txt
├── docs/
├── tests/
├── demo/
└── README.md
```

Dataset Notes:

The dataset can be any set of semantically related documents from domains such as legal case orders, technical reports, business documents, medical research, or policy papers.

Free Deployment Platforms:

- Render
- Railway
- Replit
- Hugging Face Spaces
- Vercel

Contact

For any questions during the internship or to submit your deliverables, please reach out to:

Divyansh Sharma – Wasserstoff
Email: divyansh.sharma@thewasserstoff.com