

Kalpit Pvt Ltd, UK - AI Intern Hiring : Assignment 2

The Task

Now that you have built a functional RAG prototype, your next task is to implement a comprehensive evaluation framework to measure its performance across multiple documents. You will create a testing system that quantitatively assesses the quality of your Q&A system using standard NLP metrics and comparative analysis.

We are providing:

- Document Corpus of Dr. Ambedkar's works
 - Test dataset with 25 pre-defined Q&A pairs
 - Your job is to implement evaluation metrics and analyze performance
-

Document Corpus - See Attached file

Test Dataset - See Attached file

Your Evaluation Tasks

1. Implement Comprehensive Evaluation Metrics

Your system must implement following metrics:

Retrieval Metrics:

- Hit Rate
- Mean Reciprocal Rank (MRR)
- Precision@K

Answer Quality Metrics:

- Answer Relevance
- Faithfulness (Factuality)
- ROUGE-L Score

Semantic Metrics:

- Cosine Similarity
- BLEU Score

2. Comparative Chunking Analysis

Test and compare 3 different chunking strategies:

- Small chunks: 200-300 characters
- Medium chunks: 500-600 characters
- Large chunks: 800-1000 characters

3. Performance Analysis

- Measure retrieval accuracy across all 25 questions
 - Analyze answer quality vs. chunk size
 - Identify common failure modes
 - Recommend optimal configuration
-

Technical Requirements

Required Deliverables:

1. evaluation.py - Main evaluation script with all metrics
2. test_results.json - Output of your evaluation runs
3. corpus/ folder containing all 6 documents

4. test_dataset.json - The provided test dataset
5. results_analysis.md - Detailed findings and recommendations
6. Updated requirements.txt with evaluation dependencies
7. Updated README.md with evaluation instructions

Technical Stack:

- Python 3.8+ with LangChain framework
- ChromaDB vector store
- HuggingFaceEmbeddings (sentence-transformers/all-MiniLM-L6-v2)
- Ollama with Mistral 7B
- Additional: ragas, rouge-score, nltk, scikit-learn for evaluation

Implementation Approach:

1. Start with retrieval metrics - Test if correct documents are found
2. Add answer quality metrics - Evaluate generated answers
3. Run comparative analysis - Test different chunking strategies
4. Analyze results - Identify patterns and recommendations

Submission Instructions

1. Update your existing GitHub repository AmbedkarGPT-Intern-Task
2. Add all new deliverables in the repository
3. Ensure your evaluation code runs without errors
4. Share the updated repository URL via our portal/email
5. You have 4 days to complete this assignment

Expected Output

Your evaluation should provide clear answers to:

- Which chunking strategy works best for our corpus?
- What is our system's current accuracy score?
- What are the most common failure types?
- What specific improvements would boost performance?

Good luck with your implementation!

Hiring Manager
Kalpit Pvt Ltd, UK
kalpiksingh2005@gmail.com