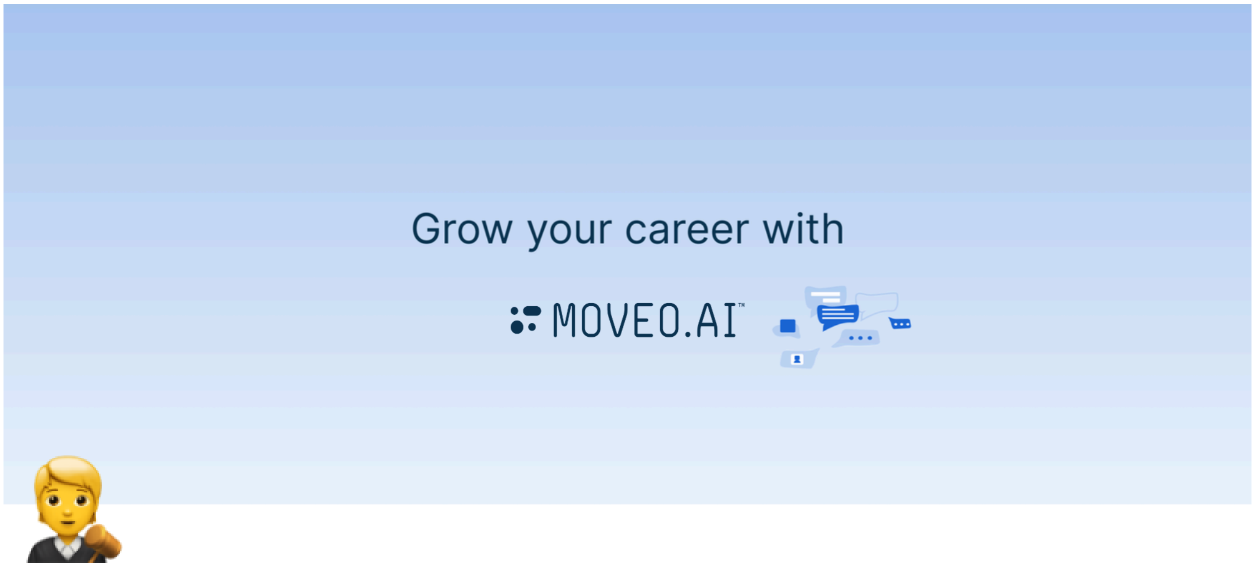

Take-Home Assignment · NLP Engineer 07/2025

Build an Automated “Judge” for RAG Answers

Overview

The task is to build an **automated evaluator (“Judge”)** that inspects the assistant answers in `rag_evaluation_07_2025.csv` and assigns quality scores.

You decide **which evaluation dimensions matter**, how to measure them, and how to combine them into an overall judgment.

Objectives (high-level)

1. **Invent your own evaluation dimensions**

   Define and justify the aspects of quality you think are important for a production RAG assistant.

2. **Implement the Judge**

   *Write **Python** code that loads the CSV, evaluates each answer along your dimensions, and produces both per-row scores and aggregate statistics. Feel* *free to use any suitable open-source libraries when you implement the judge.*

3. **Models/Algorithms**

   Your Judge may rely on LLM(s) or other algorithms. As the LLM provider, Mistral's free tier (https://auth.mistral.ai) is a zero-cost option if you need one. ***Feel free*** *to use other providers.*

4. **Report the findings**

   Generate a human-friendly report in Markdown that highlights strengths, weaknesses, and notable failure cases in the dataset.

Dataset

`rag_evaluation_07_2025.csv` can be found here

Columns:

Column	Description
`Current User Question`	Latest user utterance
`Conversation History`	Prior turns (newline-delimited)
`Fragment Texts`	Passages retrieved by the retriever
`Assistant Answer`	Assistant's response to be judged

Mandatory requirements

#	Requirement
1	**Dimension schema**: In the README, provide details that list each dimension, its possible score range, and how it contributes to the final composite score.
2	**Automated scoring**: Implement code that assigns scores *without human intervention*. Every row in the CSV must receive a value for every dimension.
3	**CLI / notebook entry point**: Running `python main.py --csv rag_evaluation_07_2025.csv` must execute the full evaluation and write a timestamped report to `reports/`. At a minimum, export the graded CSV along with aggregate statistics for the dataset.

#	Requirement
4	**Determinism controls**: Expose temperature and random-seed flags so the Judge can run deterministically.
5	**Documentation**: Include clear setup and usage instructions, plus the rationale behind your chosen dimensions and metrics.

Deliverables

Item	Where / Format
Code & data	Public **GitHub repo**
README	Setup, usage, model-switch guide, description of your dimensions
Report sample	Committed in `reports/`

Submission

- **Deadline**: within **7 calendar days** of receiving the assignment.
- Email the GitHub link (should be a public repo) to: **`nlp-team@moveo.ai`** with subject **"NLP take-home - <Your Name>"**.

Good luck - we're excited to see how you design your Judge!