# Determenistic Candidate evaluation with RAG

OMAR ABOELFETOUH

# Core Solution

## LLM IS NOT DETERMINISTIC BY NATURE - THAT'S HOW WE GONING TO TACKLE IT

- Context
  - Candidate CV
  - Job Requirements
- System Prompt
  - Evaluation matrics
  - Reasoning system prompt
  - Best practics
- LLM configuration
  - model: openAI's gpt-o4-mini
  - temprature: 0
  - top_p: 0
  - seed: 42 (fixed seed)

# RAG - Candidate Matching to a Job

01 **PARSE JOBS**

extract jobs only form the data, parse each json on an approperiate string, ensuring comprehension, and no context loss

02 **VECTORIZE JOBS**

Save the parsed jobs into a vector database (FIASS)

03 **CHOOSE RANDOM CANDIDATE**

From the data, fetch a random candidate, parse it into an approperiate string, embedd the string with the same embedding model

03 **RETRIEVAL**

using FIASS similarity search to fetch the top 3-5 jobs matches this candidate

# Parsing Techniques

01

**TECHNIQUE SELECTION (CUSTOM TEXT-CLEANING AND JSON PARSING APPROACH)**

designed to reliably extract machine-readable JSON from free-form LLM output.

02

**WHY WE CHOSE THE CURRENT PARSING TECHNIQUE?**

Because it's simple, deterministic, and fast, ensuring consistent JSON extraction under fixed model conditions.
It avoids unnecessary complexity since our LLM already outputs well-structured JSON at temperature 0.
This makes the pipeline stable, transparent, and easy to debug for large-scale evaluations.

# Similarity Search

## WHAT IS THE THRESHOLD ?

The threshold defines the minimum cosine similarity score (e.g., 0.75 or 0.90) required for a job to be considered a relevant match in FAISS retrieval.
In this project, 0.90 marks excellent semantic alignment, ensuring only strong, contextually similar jobs are passed to the LLM evaluator.

## WHY ?

We set it empirically to balance recall and precision — high enough to filter weak matches but not so strict that it excludes valid ones.

# LLM Block

**EXPERIMENTATION AND RATIONALE OVERVIEW**

The problem was that LLMs are inherently non-deterministic — the same input can yield slightly different outputs due to probabilistic token sampling.

- To ensure consistent and repeatable scoring, we enforced a deterministic setup:
- Temperature = 0 to remove randomness in generation.
- Top-p = 0.5 to limit sampling diversity.
- Chosen models: we tested GPT-4o-mini and Gemini 2.0 Flash, selecting the one with the lowest output variance.
- System prompt: we broke down the evaluation into clear, step-by-step criteria (skills, experience, location, culture) and enforced strict JSON formatting, minimizing hallucinations and making the model's reasoning stable and explainable.

# LLM Block

**THE PROBLEM WAS THAT LLMS ARE INHERENTLY NON-DETERMINISTIC, THE SAME INPUT CAN YIELD SLIGHTLY DIFFERENT OUTPUTS DUE TO PROBABILISTIC TOKEN SAMPLING. TO ENSURE CONSISTENT AND REPEATABLE SCORING, WE ENFORCED A DETERMINISTIC SETUP:**

**01 TEMPERATURE**

= 0 to remove randomness in generation.

**02 TOP-P**

= 0.5 to limit sampling diversity.

**03 CHOSEN MODELS**

we tested GPT-4o-mini and Gemini 2.0 Flash, selecting the one with the lowest output variance.

**04 SYSTEM PROMPT**

we broke down the evaluation into clear, step-by-step criteria (skills, experience, location, culture) and enforced strict JSON formatting, minimizing hallucinations and making the model's reasoning stable and explainable.

# Results - LLM consistency test

```
🤖 Initializing OpenAI LLM...

🔄 Running 10 evaluations...
   Run 1/10... Score: 71.4
   Run 2/10... Score: 71.4
   Run 3/10... Score: 71.4
   Run 4/10... Score: 71.4
   Run 5/10... Score: 71.4
   Run 6/10... Score: 71.4
   Run 7/10... Score: 71.4
   Run 8/10... Score: 71.4
   Run 9/10... Score: 71.4
   Run 10/10... Score: 71.4

--------------------------------------------------------------------------------
INDIVIDUAL RESULTS:
--------------------------------------------------------------------------------
Run 1: Score = 71.4, Recommendation = None
Run 2: Score = 71.4, Recommendation = None
Run 3: Score = 71.4, Recommendation = None
Run 4: Score = 71.4, Recommendation = None
Run 5: Score = 71.4, Recommendation = None
Run 6: Score = 71.4, Recommendation = None
Run 7: Score = 71.4, Recommendation = None
Run 8: Score = 71.4, Recommendation = None
Run 9: Score = 71.4, Recommendation = None
Run 10: Score = 71.4, Recommendation = None

================================================================================
DETAILED CONSISTENCY ANALYSIS
================================================================================

📊 SCORE STATISTICS:
   Runs:              10/10 successful
   Score Range:       71.4% - 71.4%
   Average:           71.40%
   Median:            71.4%
   Mode:              71.4% (10 times)
   Std Deviation:     0.00
   Consistency Score: 100/100
```

# Results - LLM consistency test

```
================================================================================

🏆 Rank 1: Similarity Score: 0.7759
   Job ID: JOB-000184
   Title: QA Engineer
   Company: Orlion Analytics
   Location: Alexandria, Egypt
   Sector: Software Engineering
   Requirements: Proficiency in Spring Boot...
--------------------------------------------------------------------------------

🏆 Rank 2: Similarity Score: 0.7661
   Job ID: JOB-000368
   Title: QA Engineer
   Company: Arcadia Systems
   Location: Cairo, Egypt
   Sector: Software Engineering
   Requirements: Proficiency in MongoDB...
--------------------------------------------------------------------------------

🏆 Rank 3: Similarity Score: 0.7475
   Job ID: JOB-000417
   Title: QA Engineer
   Company: Crestel Technologies
   Location: Alexandria, Egypt
   Sector: Software Engineering
   Requirements: Proficiency in MongoDB...
--------------------------------------------------------------------------------

🏆 Rank 4: Similarity Score: 0.6695
   Job ID: JOB-000060
   Title: QA Engineer
   Company: Velora Technologies
   Location: Remote – EMEA
   Sector: Software Engineering
   Requirements: Proficiency in TypeScript...
--------------------------------------------------------------------------------

🏆 Rank 5: Similarity Score: 0.6659
   Job ID: JOB-000059
   Title: QA Engineer
   Company: Orlion Solutions
   Location: Remote – EMEA
   Sector: Software Engineering
   Requirements: Proficiency in CI/CD...
--------------------------------------------------------------------------------
```

# Notes

- The solution mainly focuses on achieving deterministic behavior and maintaining a very narrow margin of error without relying on caching.
- We can experiment with different LLMs to observe and compare their behaviors.
- Fine-tuning can further improve the solution by training the model to calculate evaluation metrics more accurately.
- The solution does not emphasize any evaluation metrics or the accuracy of job matching; its main focus is to achieve consistent results across different LLM calls.
- Once deterministic behavior is achieved, we can align and fine-tune the metrics to reflect the actual business evaluation criteria.