PS-4: Retrieval Augmented Generation based Question and Answering System

1. PS Description:

Today, a lot of information is stored in various unstructured documents. Often, the information is stored across multiple documents. To sift through humongous documents and extract vital information to serve user queries is crucial and challenging.

The subject challenge '**Retrieval Augmented Generation based Question and Answering System**' entails development of an information correlation solution across multiple documents to generate summaries that are grounded in evidence.

The RAG solution should combine semantic retrieval and generative reasoning.

2. Objectives:

1) Relevant documents retrieval from a diverse documents corpus.
2) Generate fluent, factually grounded and explainable natural language responses.
3) Minimize hallucination & ensure all claims are traceable to source content

3. Stages, Datasets, Desired Outputs & Rules:

Stage	Timeline	Datasets	Desired Outputs	Metrics
I (03 months in total)	T0+90 days	The dataset will be selected from the following sources (but not limited to): 1. Cord-19 2. Wikihop 3. Hotpot QA 4. EU-Legislation Corpus (Only English docs) 5. APT NOTES 6. RobustQA 7. MultiHop-QA	In the first stage the participants would be judged on their retrieval algorithm	**Retrieval Combined Score based on the below metrics** 1. Precison@k (20% weightage) 2. Recall@k (50%) 3. NDCG (30%)
II	04 Months	The dataset will be selected from the	In the second stage the participants would be judged on their retrieval	**Combined Score – Retrieval 65%**

		following sources (but not limited to): 1. Cord-19 2. Wikihop 3. Hotpot QA 4. EU-Legislation Corpus (Only English docs) 5. APT NOTES 6. RobustQA 7. MultiHop-QA	algorithm and the generation algorithm	**Generation 35% with score of retrieval and generation as follows: -** **Retrieval– Combined Score based on the below metrics** 1. Precison@k (20% weightage) 2. Recall@k (50%) 3. NDCG (30%) **Generation – Combined Score based on the below metrics** 1. Rouge(25% weightage) 2. Meteor(15% weightage) 3. Hallucination Rate (inverse) – 60% weightage
III	05 months	The dataset will be from the organization specific documents and will be given toStage-3 participants in our incubation centre	In the third stage the participants would be judged on their retrieval algorithm and the generation algorithm. The generation in this stage will also be multi-turn.	**Combined Score – Retrieval 65%** **Generation 35% with score of retrieval and generation as follows: -** **Retrieval– Combined Score based on the below metrics** 1. Precison@k (20% weightage) 2. Recall@k (50%)

				3. NDCG (30%) **Generation – Combined Score based on the below metrics** 1. Rouge (25% weightage) 2. Meteor (15% weightage) 3. Hallucination Rate (inverse) – 60% weightage

Note: Training data would be in English and use of online APIs should be avoided.

4. Dataset arrangements for Stage-1 & 2

a. **Training Dataset (upto 10 GB):** train_set.zip. Participants should use this dataset for solution development. It will be released at T0 i.e. 01-Aug-2025. Participants are free to supplement it with datasets of their choice.

b. **Mock Dataset (upto 5 GB):** mock_set.zip. Participant should test their solution on this dataset, but will not have access to the corresponding ground truth during the Challenge. This dataset is for self-assessment and will not be used for evaluation for selection. Participants need to submit the results of their solution on this set. This set will be released on T0 + 45 days i.e. 15-Sep-2025 (single dataset consisting of files and queries in csv format). A leaderboard will be published on this Mock Dataset.

c. **Shortlisting Dataset (upto 5 GB):** short_listing_set.zip will be released on at 1100h on 03 Nov 2025. Based on the results submitted on this dataset, by 2359h on 04 Nov 2025, 15-20 participants will be shortlisted for offline solution evaluation.

d. **Holdout Dataset:** After the Challenge deadline, a private ranking will be computed using this holdout set. This set will be made available during final offline evaluation for which details would be communicated to the 15-20 shortlisted participants..

e. **All the dataset would be named as <xyz>.<file_format>**

5. **Input/ Output Instructions**

   a. Input

The input to the participants would be a csv file consisting of queries with serial numbers.

c. Output

The participants need to submit one json corresponding to each query. The name of the response would be <query_number>.json. The json should have the following format:

{

“query”: <actual text query>,

“response”: [<xyz1.file_format>,<xyz2.file_format>, …]

}

All json files be zipped and submitted as <startup name>_PS4.zip

6. **Online solution evaluation during Stage-1 (Mock Datasets)**
   - **a.** Solutions are expected to be submitted on Thursday from week commencing 15 Sep 2025, on the Mock Dataset.
   - **b.** Leadersboard will be updated every Tuesday**.**
   - **c.** Scores will be computed based on the evaluation metrics as under: -

Category	**Criteria**	**Description**	**%Weight**
Metric Evaluation	**Retrieval– Combined Score based on the below metrics** 1. Precison@k (20% weightage) 2. Recall@k (50%) 3. NDCG (30%)	Score based on test dataset	100

7. **<u>Selection of 15-20 participants for offline-evaluation in Stage-1</u>**
   - a. On 3rd Nov 2025, the details of Shortlisting Dataset will be made available on the website. Results generated on the Shortlisting Dataset will be evaluated for final selection of 15-20 participants. The number

may vary based on the overall performance at the discretion of the Jury for this Problem Statement. Submissions found Incomplete in any manner will not be considered for further processing. The shortlisted participants will be published alongwith the cutoff score as per the evaluation criteria. Participants individual scores will be shared over the email.

b. Any kind of unfair means be avoided while developing and generating the solution and results, failing which will leads to cancellation of participation for the grand challenge and organisers can call the next participant from leader-board for evaluation.

c. Scores will be computed based on the evaluation metrics indicated below:

Category	**Criteria**	**Description**	**%Weight**
Metric Evaluation	**Retrieval– Combined Score based on the below metrics** 4. Precison@k (20% weightage) 5. Recall@k (50%) 6. NDCG (30%)	Score based on test dataset	100

d. Based on performance on the shortlisting datasets top 15-20 participants will be called for an offline evaluation. Scores will be computed based on evaluation metric as above.

8. **Solution Evaluation at the end of Stage-1 Deadline (Holdout Dataset)**

a. Shortlisted participants will be asked to demonstrate their solution at IIT Delhi on completion of stage-1 deadline.

b. Participants will be allotted slots in which they need to run their solution on reference data provided by the organizers on given resources with following specifications: -

i. OS – Ubuntu 24.04 LTS
ii. CPU – 48+ core
iii. RAM – 256+ GB
iv. GPU - A-100, 40/80 GB
v. Solution Demo Duration: 02 Hours for each selected participant

c. Based on the results from solution demonstration and presentation, final scores will be computed based on Evaluation Metrics as mentioned below:

Category	**Criteria**	**Description**	**% Weight**
Solution Evaluation	**Combined Score – Retrieval 65%** **Generation 35% with score of retrieval and generation as follows:-** **Retrieval– Combined Score based on the below metrics** 1. Precison@k (20% weightage) 2. Recall@k (50%) 3. NDCG (30%) **Generation – Combined Score based on the below metrics** 1. Rouge(25% weightage) 2. Meteor(15% weightage) Hallucination Rate (inverse) – 60% weightage	**Score based on official metric on hidden hold-out test dataset**	45
Robustness	Efficiency	Solution Execution time on hold-out test dataset	10
Resource Utilization	Solution Memory Footprint	Memory used by Solution during execution	10

Approach	Methodologies of Solution Development	Start-up need to present Solution development approaches & proposed Architecture	25
Team Capabilities	Technical Capabilities of Start-up Team	Team Composition, Qualifications, Experience and ability to complete the challenge end to end.	10

d. Participants are free to use any language or development framework for the solution.
e. At most top 6 teams will be selected based on final score for Phase-2

7. Evaluation Criteria for Stage-II and Stage-III would be similar as above. It would be released to concerned participants before the beginning of Stage-2 and Stage-3 respectively.

8. **Sessions with Mentors\Experts**
   a. For Stage-1, the organisers plan to meet participants via online meet or email to resolve their doubts, if any. This provision will be made active from 15th Aug 2025 and details regarding interaction will be shared on this website. Kindly keep viewing this website regularly for updates on this.
   b. There will be sessions with Mentors\Experts in Stage-2 and Stage-3 for the willing selected participants to help them in achieving the best solutions.