Microsoft

SOCAR

BAKI ALİ NEFT MƏKTƏBİ
BAKU HIGHER OIL SCHOOL

SOCAR Hackathon 2025

Main topic and details

Hackathon 2025

AI track

Details of AI track

Data engineering Track

Details of Data Engineering track

Artificial Intelligence Track

Develop a complete solution that transforms historical handwritten and printed documents into an interactive, searchable knowledge base accessible through an intelligent chat agent for upstream use cases.

Definition of Input

Azerbaijani Latin, Typewritten

Easiest documents with better readability and easier font

Azerbaijani Cyrilic, Typewritten

Medium Difficulty documents with moderate readability like real scan documents

Azerbaijani Latin, handwritten

The most difficult dataset, with handwritten information

Practical exercise

Evaluation

OCR evaluation can give max 500 points

Evaluation

Evaluation of chatbot can give max 300 points

OCR of PDFs

OCR of given different types of documents

Chatbot

Creation of chatbot based on give documents

Presentation

Demonstrate all your research 200 Points

Scoring of OCR

Teams should create OCR endpoint in which input will be PDF and output is specific format given in guideline. For compatibility check, space in SOCAR Huggingface account can be used.

CSR = (1 – CER)*100*coef
WSR = (1 – WER)*50*coef

	Easy	Medium	Hard	Total
CSR	87,5	112,5	137,5	337,5
WSR	37,5	56,25	68,75	162,5
Total	125	168,75	206,25	500

Scoring of Chatbot

Teams should create Chatbot endpoint in which input will be question and output is specific format given in guideline. For compatibility check, space in SOCAR Huggingface account can be used.

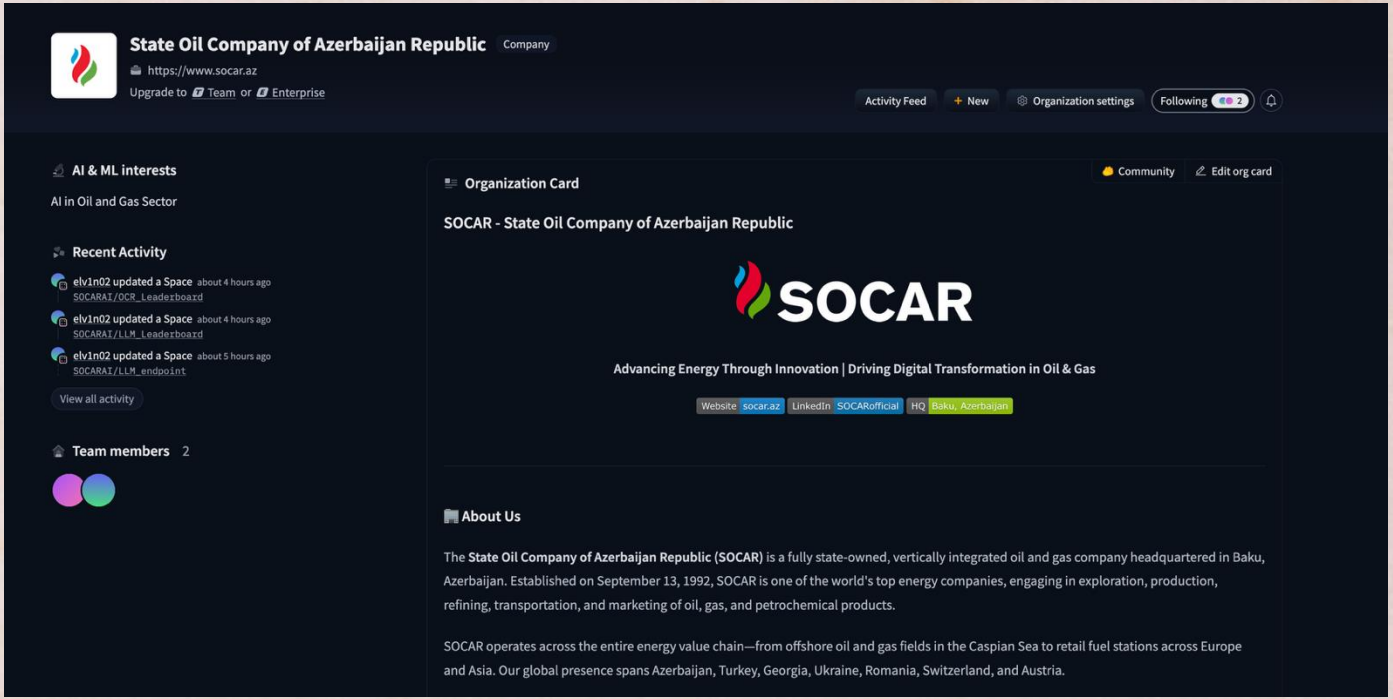
Number of questions

	Easy	Medium	Hard	Total
Printed	2	2	2	6
Cyrillic Printed	2	2	3	7
Handwriting	2	3	2	7
Total	6	7	7	20

Scoring of questions

	Easy	Medium	Hard	Total
Printed	6,45	9,675	12,9	58,05
Cyrillic Printed	9,675	14,55	19,35	106,5
Handwriting	12,9	19,35	25,8	135,45
Total	58,05	106,5	135,45	300

Huggingface space



Data Engineering Track

A high-intensity data engineering challenge where teams reconstruct complex data and transform it into actionable insights.

Task description

Step 01
Decode data

Step 02
Create data models

Step 03
Work on data marts

Step 04
Set up environement

Step 05
Run the ETL job

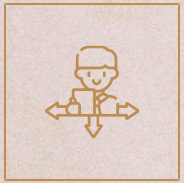
Given Data

Column Name	Definition
well_id	Unique identifier for the well.
well_name	Human-readable name of the well.
location_lat	Latitude coordinate of the well's location.
location_long	Longitude coordinate of the well's location.
operator	Company or entity responsible for operating the well.
spud_date	Date when drilling of the well officially began.
sensor_type	Type or category of the sensor (e.g., pressure, vibration).
calibration_date	Date when the sensor was last calibrated.
survey_type	Name of the survey type (e.g., Acoustic_2D, Acoustic_3D, Acoustic_VSP).
depth_ft	Depth measurement in feet at which the reading was taken.
amplitude	Measured amplitude value.
quality_flag	Indicator of data quality (e.g., valid, suspect, invalid).

Submissions

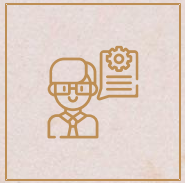
README.md

Have a very detailed README.md file in your github repository like a White paper of your project.



Solutions folder

Store scripts for task solutions in the given folder on github. It is necessary for evaluation.



Presentation

Prepare a presentation for your project showcasing what you have done.

Evaluation

Criteria	Weight
Presentation	10
Suggested architecture	15
Computational efficiency	10
Tasks	65

Tips:

- Push to GitHub frequently.
- Try using architectural diagrams.
- Pay attention to documentation.
- Do not get stuck in one problem for too long.
- We are assessing your technical skills not geological, so do not dive deeply into the data, focus on demonstrating technical complexity to the jury members.

Microsoft

SOCAR

BAKI ALİ NEFT MƏKTƏBİ
BAKU HIGHER OIL SCHOOL

Thanks