

## Bajaj Finserv Health Limited – DATATHON

(Where Individual Brilliance Meets Data)

**Start Date:** 28th Nov'25 | 12:00 PM

**End Date:** 30th Nov'25 | 12:00 PM

**Opportunity:** Final Placement Offer

**Eligibility:** Class of 2026 (**Final Year Students**) | **Format:** Individual Datathon

**Degree:** B.Tech & B. Tech + M.Tech (**Integrated**) | **All streams are eligible for the Datathon**

**Registration Link :** [https://www.surveymonkey.com/r/datathon\\_25](https://www.surveymonkey.com/r/datathon_25)

### About Bajaj Finserv Health | Background and Purpose:

Bajaj Finserv Health and affiliate companies are in the business of Health Insurance Claims processing OPD (out-patient) & IPD (in-patient). We process millions of bills (Invoices) a month and have to ensure we 'pay right'.

These bills are received as scanned photos and pdfs (running into 30-40 pages). Bills have to be reviewed by processing agents and appropriate amounts have to be extracted to be entered into the system. We are digitizing/automating this process of bill data extraction – extracting the line-item data and bill totals. The problem statement below is in-line with "**Improving' the accuracy of the bill extraction process**".

### Problem statement:

You will be given a set of bills (invoices) with multiple pages (training data); you need to design and develop data extraction/summarization models that will 'extract the line item details from these bills and also provide 'individual line item amount', 'Sub-total' and 'Final Total' of the bill amounts extracted (sub-totals will be required only where they exist in the bill). 'Final Total' will be sum (of all individual line items in the bills) without double-counting.

It's important to ensure that, you **don't miss out on any line-item** entries and at the same time **don't double count any entries**. The closer your 'Total AI extracted amounts' is equal to the 'Actual Bill Total' -the better your accuracy.

### Tools:

You are free to use any LLM models or services (like OCR) available in the market. Programming languages such as Python, Java can be used to create APIs.

**(Any cost incurred for the usage of any paid tools should be borne by the students)**

### Input provided (to all teams)

Training data consisting of bills with varying degrees of formats (simple to more complex/multilingual/handwritten etc.); A sample of outputs for these training data will also be provided

### Deliverable:

- Models that perform these extractions with high accuracy (to be submitted in a private Github repo)
- An API that will accept a document(s) with multiple pages and provide output in expected JSON format. [Link to Download the JSON Format](#).
- This API will be used by our teams to score your models and rank you on the leaderboard; any updates to the problem statement and the API structure will be published on the below link. <https://datathon.healthrx.co.in/#problem-statement>

- A pitch deck – that represents your overall architecture and ‘tech/model stack < 2-3 pages>; Call out your differentiators. [Link to Download the Pitch deck Format.](#)

## Differentiators:

- For some of the documents we provide you, you will require pre-processing to improve ‘extraction accuracy’, please call out any ‘pre-processing’ you have used
- Some of the documents we provide will also have elements of ‘fraud’, e.g. inconsistent ‘fonts types’, ‘overwriting with whitener’ etc. if you are able to catch these – this is also your differentiator

## Criteria for Interview Selection

- Overall score in the leader board
- Pitch deck and architecture used
- Differentiators
- Latency of APIs
- Github code review

**\*Its important that everyone solves this assignment individually; Anyone we suspect to have plagiarized the solutions will be automatically disqualified \***

## Appendix 1:

Illustration of sample documents we will provide for training and illustration of output from these documents (we will give the exact json structure you need to create)

Registration No. : GUR/ LT-BU/H/2025/000001		122001, Gurugram - 122001																									
		0124-4939680 / 681 / 9717189900, 910351443																									
		avisehospital@gmail.com																									
 <b>AVISE HOSPITAL</b> <b>SUPERSPECIALITY</b> (A UNIT OF SIDHHI HEALTHCARE PRIVATE LIMITED)  www.avisehospital.com																											
<b>PATIENT BILL</b>																											
Bill No. : CA/0000569		CREDIT	Bill Date : 10-Sep-2025																								
ROOM / BED NO.	GENERAL WARD / GW-04	D.O.A.	09/Sep/2025 12:33:28 PM																								
NO. OF DAYS	2	D.O.D.	10/Sep/2025 1:35:43 PM																								
NAME OF PATIENT		IPD No.	20250867 UHID   111111																								
ADDRESS		Age/Sex	36/ Male																								
MOBILE No.	NAME OF CONSULTANT																										
TPA NAME	CORPORATE NAME																										
INSURANCE COMPANY NAME	POLICY NO.																										
Sr.No.	Particulars	Qty.	Rate																								
			Amt.																								
			Disc.																								
			Amount																								
1. 1	WARD SERVICES ROOMBED CHARGES Ward/ Room/Bed No. GENERAL WARD/GW-04	2.00	1500.00																								
			3000.00																								
			3000.00																								
2. 1	LAB CHARGES LFT	1.00	950.00																								
2. 2	HBSAG	1.00	500.00																								
2. 3	CBC	1.00	500.00																								
2. 4	HIV	1.00	950.00																								
2. 5	KFT	1.00	1800.00																								
2. 6	HCV	1.00	500.00																								
2. 7	PTINR	1.00	500.00																								
3.	GYNAECOLOGY ANESTHESIA	1.00	20000.00																								
4.	IPD SERVICES SURGEON SPECIALIST DR. VISIT	3.00	800.00																								
4. 1	RIRS SURGERY CHARGE	1.00	40000.00																								
4. 2	OT CHARGE	1.00	20000.00																								
4. 3	MEDICINE	1.00	25705.00																								
5. 1	PHARMACY	1.00	25705.00																								
			Sub Total :																								
			116755.00																								
			Other Amt. :																								
			116755.00																								
			Net Bill Amt. :																								
			116755.00																								
			Advance Amt. :																								
			116755.00																								
Final Rcpt. Details Net Payable Amount : 116755.00																											
<table border="1"> <tr> <td colspan="4">Advance Details</td> </tr> <tr> <td>Mode</td> <td>Rept. No.</td> <td>Date</td> <td>Adv. Amt.</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> </tr> </table> <table border="1"> <tr> <td colspan="4">Final Rcpt. Details</td> </tr> <tr> <td>Mode</td> <td>Rcpt/Refund Amt.</td> <td>Receipt No.</td> <td></td> </tr> <tr> <td>CREDIT</td> <td>116755.00</td> <td></td> <td></td> </tr> </table>				Advance Details				Mode	Rept. No.	Date	Adv. Amt.					Final Rcpt. Details				Mode	Rcpt/Refund Amt.	Receipt No.		CREDIT	116755.00		
Advance Details																											
Mode	Rept. No.	Date	Adv. Amt.																								
Final Rcpt. Details																											
Mode	Rcpt/Refund Amt.	Receipt No.																									
CREDIT	116755.00																										
 Prepared By _____																											

Note: Leaderboard scoring and final assessments are internal processes. BFHL reserves full rights to modify or update them as deemed necessary.