

Travel Partner

By Aditya Unal

Choosing LLMs

- I used huggingface inference apis

- First I used **mistralai/Mistral-Small-24B-Instruct-2501**. Since this is a small rental shop and the queries do not require large number of tokens I tried this. Although it was cheap $\sim \$0.01$ for around 20 tokens, the results very highly inaccurate and sometimes truncated.

- Next I tried **mistralai/Mistral-7B-Instruct-v0.3** but it does not work well with tool calling.

- Finally I used **mistralai/Mixtral-8x7B-Instruct-v0.1**. I only allowed `max_new_tokens = 1000`, because this is a simple chatbot, no long responses are required. This saves on cost as well

Vehicle Information and Bookings Information

- Used **SQLite** because it is a very light framework. Ideal for small businesses. Can be hosted in the free tier in google cloud console. Hence, **cost-effective** .
- I assumed the shop owner will not reserve a vehicle unless they have the hard-copy of the a valid id. Hence, booking feature is not given.
- The vehicles have a unique **vehicle_id** which can be taken as RTO issued id.
- Similarly each person that books the vehicle has their **unique_id** issued as well which can be considered driver's license number. One person can only book one vehicle, to implement one vehicle per license .
- Next slide has a few rows and columns from the vehicles and bookings database. The tables are designed in such a way that the person an SQL Query can be created by the llm according to user's preference.

	vehicle_id	vehicle_name	vehicle_type	vehicle_wheel_count	vehicle_gear_count
0	SCORPIO001	Mahindra Scorpio	Car	4	6-Speed
1	RE_CLASSIC_001	Royal Enfield Classic 350	Bike	2	5-Speed

	booking_id	customer_id	vehicle_id	booking_date	start_date	end_date
0	BKG_JSON_001	CUST_J_001	SCORPIO001	2025-06-11	2025-06-15	2025-06-18
1	BKG_JSON_002	CUST_J_002	ACTIVA_001	2025-06-11	2025-06-11	2025-06-11

Eg : Do you have a Royal Enfield
Interceptor available on 23rd June

