Ismail Syed Specialist Leader EMEA - Data Science, Vector & ML Oracle

Evaluating LLM from AI Quick Actions Catalog

Referenced Documentation

https://docs.oracle.com/en-us/iaas/data-science/using/ai-quick-actions-evaluations.htm

Description

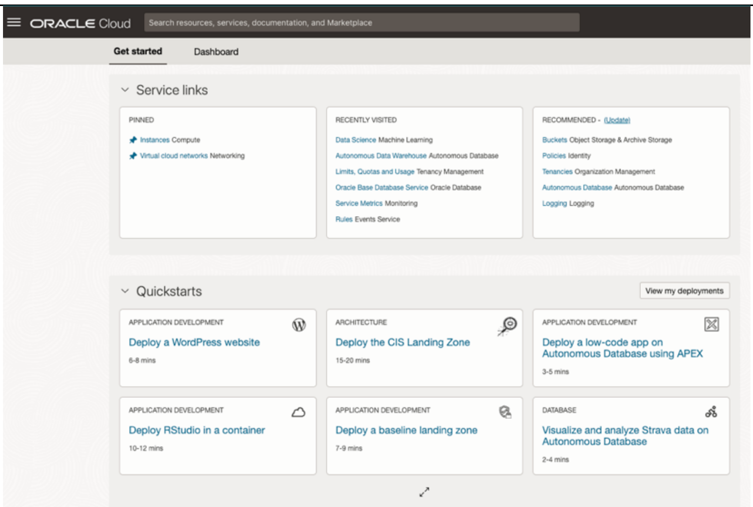
With deployed models, you can create a model evaluation to evaluate its performance. You can choose a dataset from Object Storage or upload one from the storage of the notebook you're working in. BERTScore and ROUGE are the evaluation metrics available for measuring model performance. You can save the model evaluation result in Object Storage. You can set the model evaluation parameters. Under advanced options, you can choose the compute instance shape for the evaluation and optionally enter the Stop sequence.

Pre-Requisites

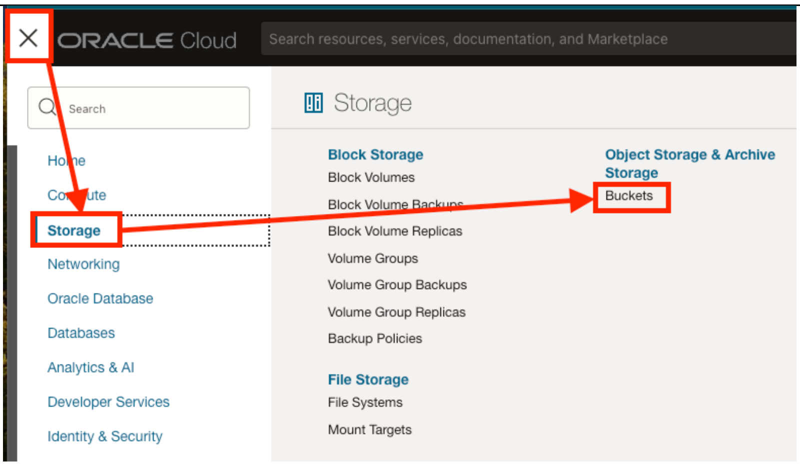
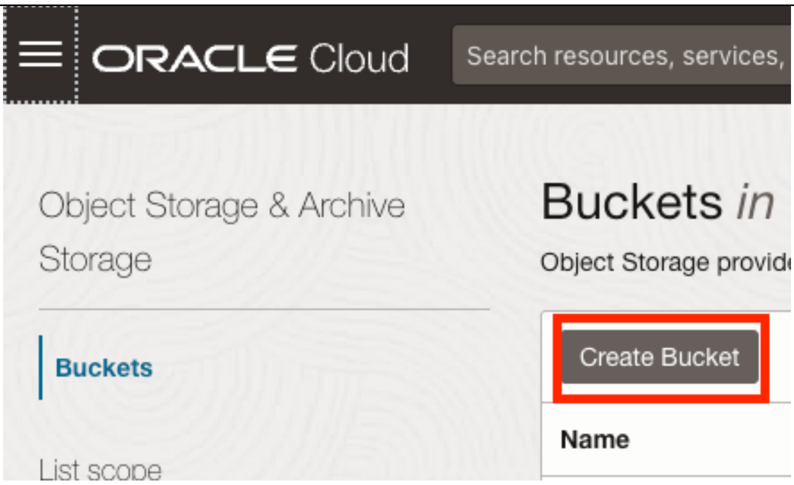
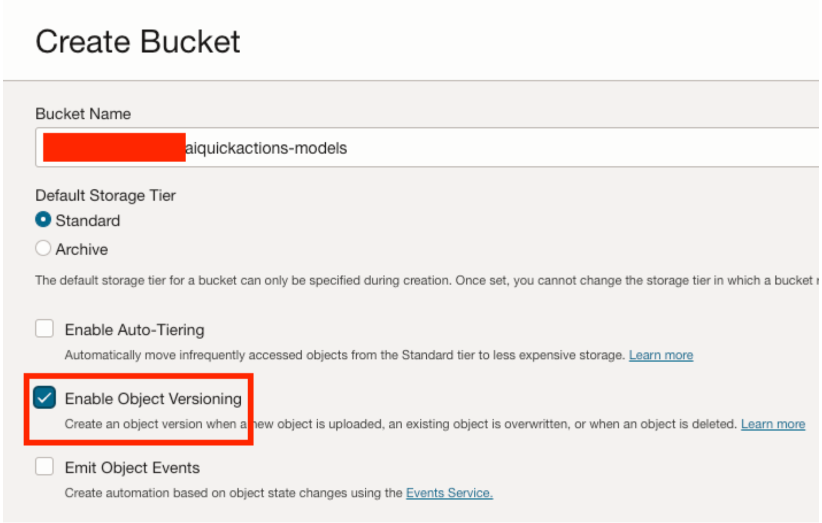
- Implement the required policies - https://docs.oracle.com/en-us/iaas/data-science/using/ai-quick-actions-set-up.htm
- Ensure you have your OCI Data Science GPU service limits raised for the GPU Shapes you plan to use. This can be done from OCI Console
- Provisioned OCI Data Science Project and Notebook Session (Must be deactivated and reactivated if created before the policies where implemented).
- Deployed AI Quick Actions Model.
- OCI Log Group & Log Created (Optional)

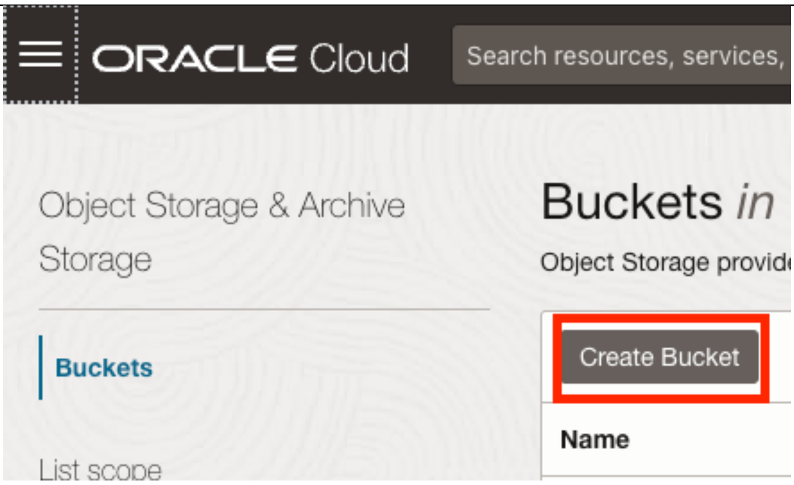
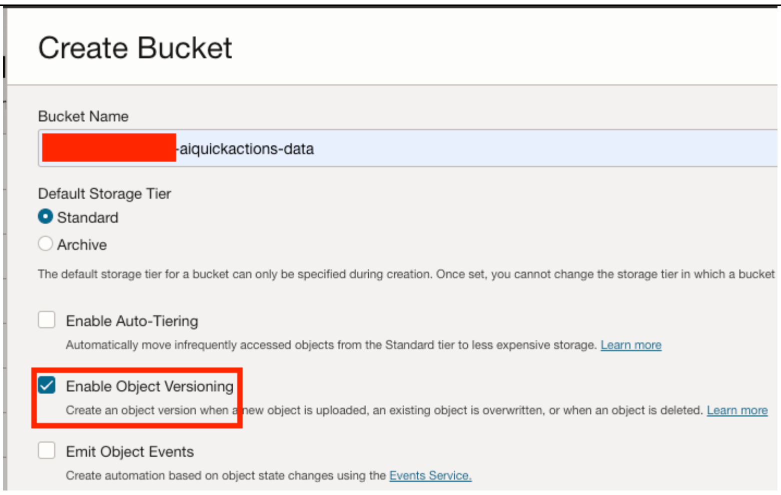
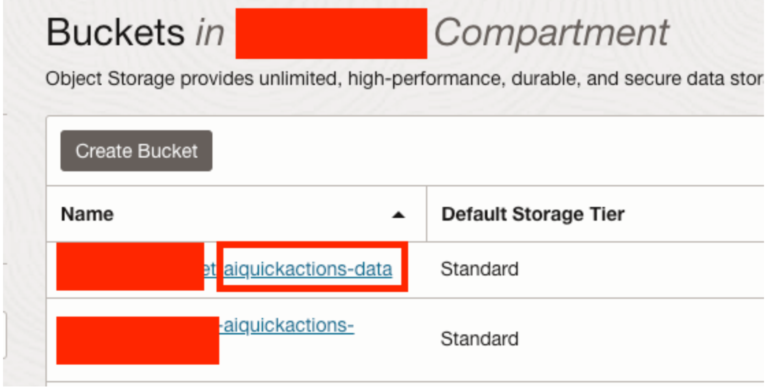
Guide

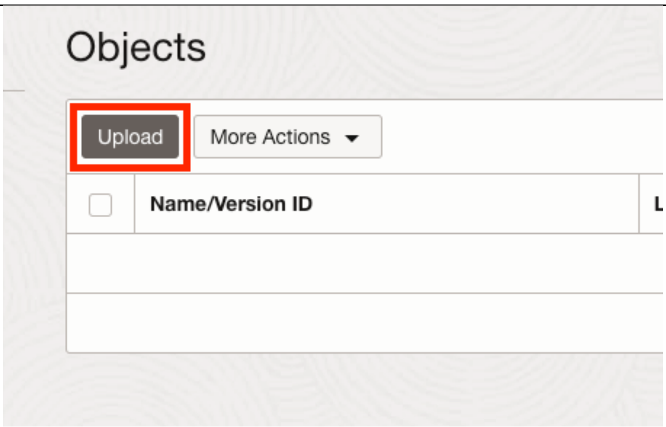
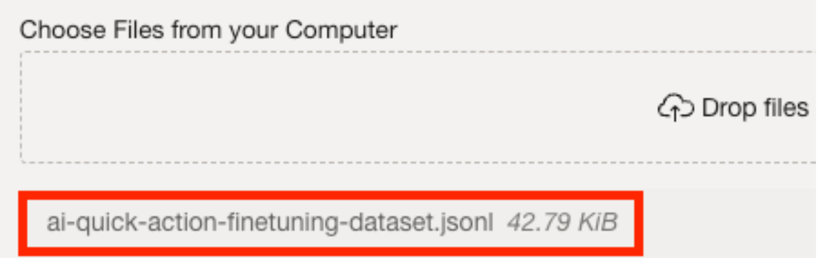
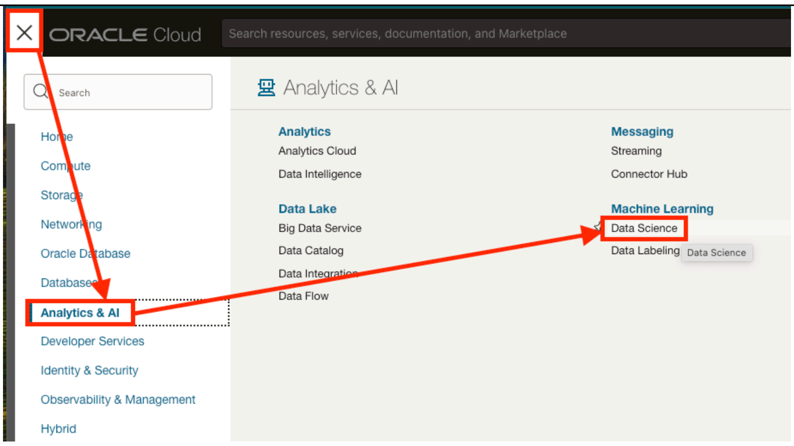
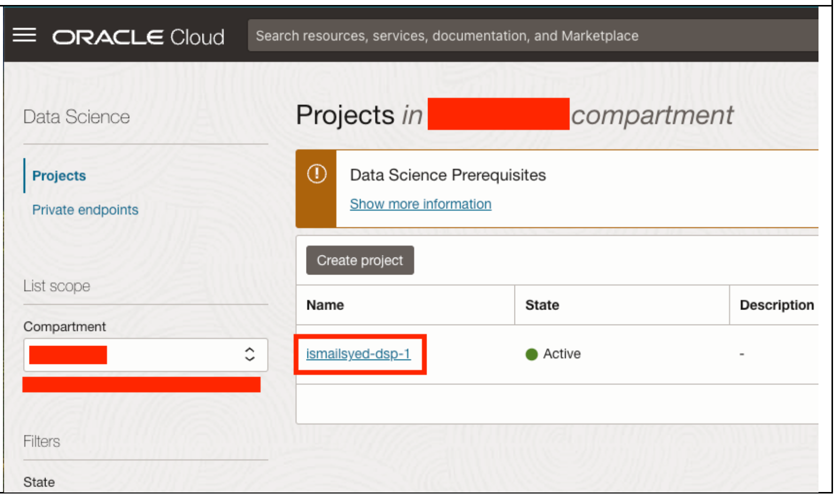
<table>
<tr><th>Step</th><th>Screenshot</th></tr>
<tr><td>Login to the Cloud Console.

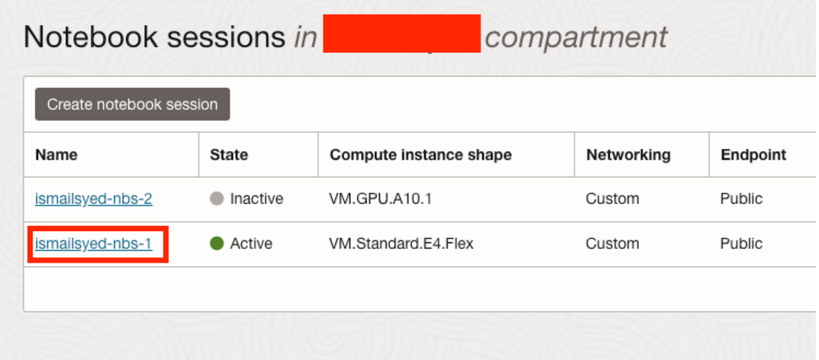
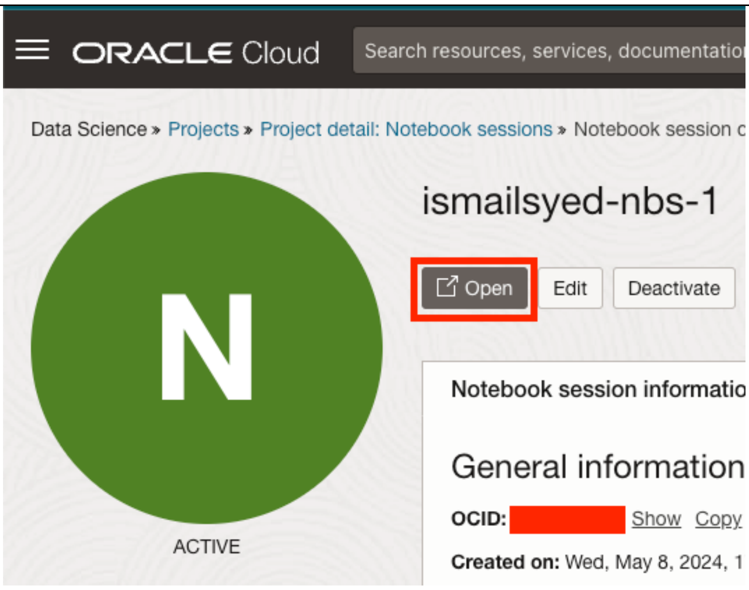
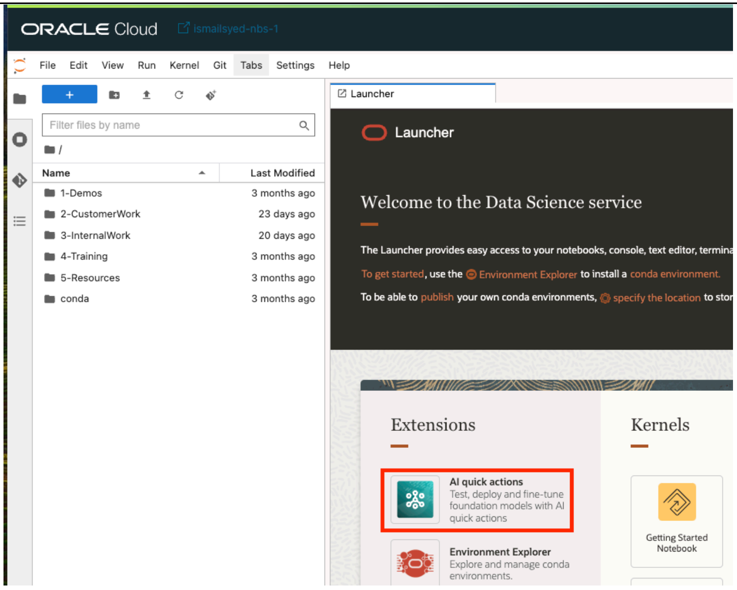
<strong><em>cloud.oracle.com</em></strong></td><td>
ORACLE Cloud
Search resources, services, documentation, and Marketplace
Get started
Dashboard
Service links
PINNED
Instances Compute
Virtual cloud networks Networking
RECENTLY VISITED
Data Science Machine Learning
Autonomous Data Warehouse Autonomous Database
Limits, Quotas and Usage Tenancy Management
Oracle Base Database Service Oracle Database
Service Metrics Monitoring
Rules Events Service
RECOMMENDED - (Update)
Buckets Object Storage & Archive Storage
Policies Identity
Tenancies Organization Management
Autonomous Database Autonomous Database
Logging Logging
Quickstarts
View my deployments
APPLICATION DEVELOPMENT
Deploy a WordPress website
6-8 mins
ARCHITECTURE
Deploy the CIS Landing Zone
15-20 mins
APPLICATION DEVELOPMENT
Deploy a low-code app on Autonomous Database using APEX
3-5 mins
APPLICATION DEVELOPMENT
Deploy RStudio in a container
10-12 mins
APPLICATION DEVELOPMENT
Deploy a baseline landing zone
7-9 mins
DATABASE
Visualize and analyze Strava data on Autonomous Database
2-4 mins

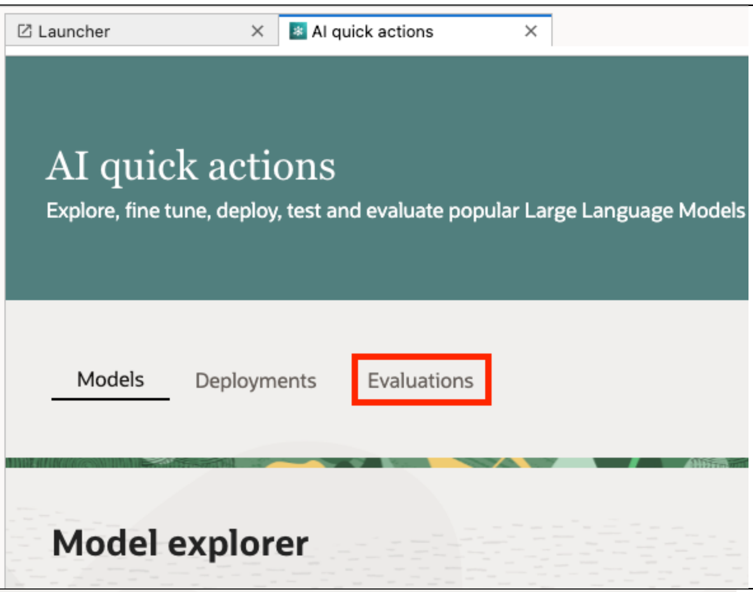
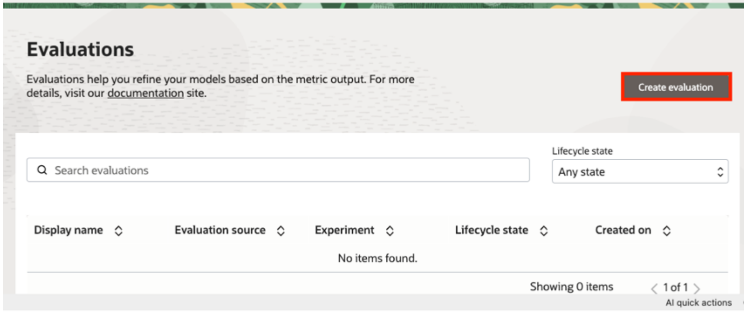
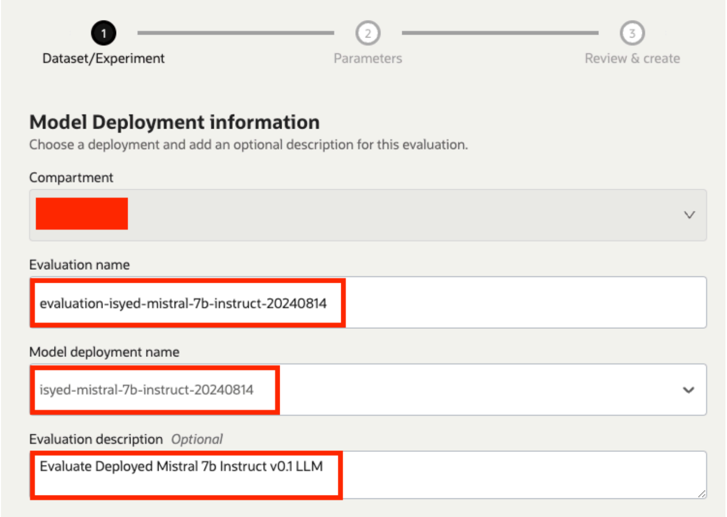
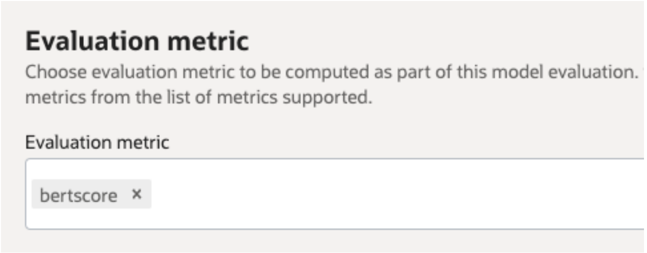
</td></tr>
</table>

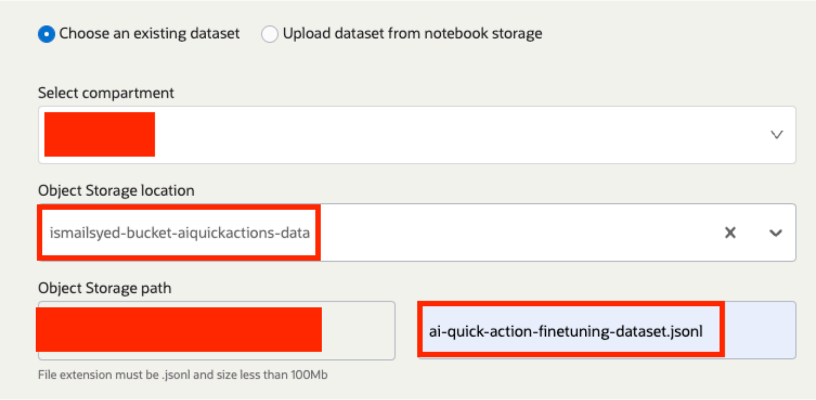
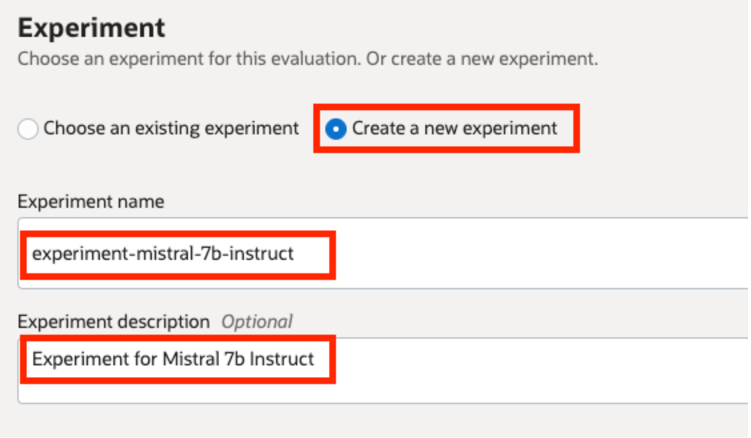
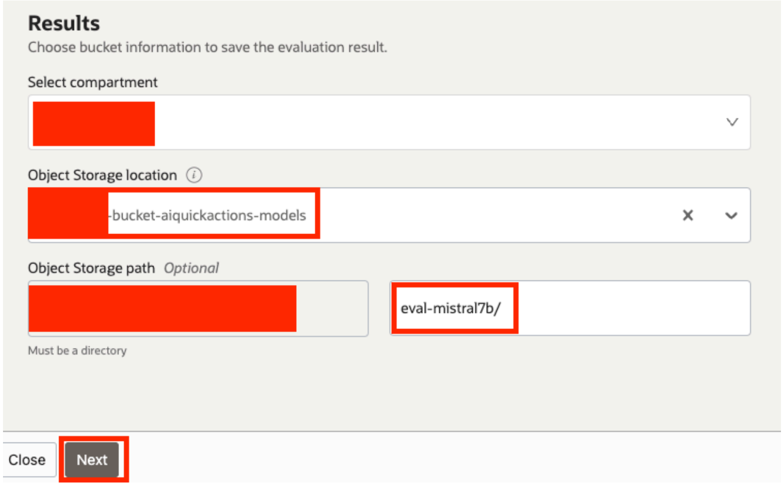
First, we will create two buckets to store our evaluation data and our evaluation results. ***Navigate to OCI Menu > Storage > Buckets.***	 ORACLE Cloud Search resources, services, documentation, and Marketplace Search Home Compute Storage Networking Oracle Database Databases Analytics & AI Developer Services Identity & Security Storage Block Storage Block Volumes Block Volume Backups Block Volume Replicas Volume Groups Volume Group Backups Volume Group Replicas Backup Policies File Storage File Systems Mount Targets Object Storage & Archive Storage Buckets
Click Create Bucket.	 ORACLE Cloud Search resources, services, Object Storage & Archive Storage Buckets List scope Buckets *in* Object Storage provid Create Bucket Name
The first bucket we create will be for our evaluation results. ***Enter a Name.*** ***Enable Object Versioning.*** ***Click Create.***	Create Bucket Bucket Name [redacted]aiquickactions-models Default Storage Tier ● Standard ○ Archive The default storage tier for a bucket can only be specified during creation. Once set, you cannot change the storage tier in which a bucket ☐ Enable Auto-Tiering Automatically move infrequently accessed objects from the Standard tier to less expensive storage. Learn more ☑ Enable Object Versioning Create an object version when a new object is uploaded, an existing object is overwritten, or when an object is deleted. Learn more ☐ Emit Object Events Create automation based on object state changes using the Events Service.

We will then create the bucket for our evaluation data. ***Click Create Bucket.***	 ORACLE Cloud Search resources, services, Object Storage & Archive Storage Buckets List scope Buckets *in* Object Storage provid Create Bucket Name
Enter a Name. ***Enable Object Versioning.*** ***Click Create.***	 Create Bucket Bucket Name [redacted]-aiquickactions-data Default Storage Tier ● Standard ○ Archive The default storage tier for a bucket can only be specified during creation. Once set, you cannot change the storage tier in which a bucket ☐ Enable Auto-Tiering Automatically move infrequently accessed objects from the Standard tier to less expensive storage. Learn more ☑ Enable Object Versioning Create an object version when a new object is uploaded, an existing object is overwritten, or when an object is deleted. Learn more ☐ Emit Object Events Create automation based on object state changes using the Events Service.
Click on our Data bucket we just created.	 Buckets *in* [redacted] *Compartment* Object Storage provides unlimited, high-performance, durable, and secure data stor Create Bucket Name — Default Storage Tier [redacted]aiquickactions-data — Standard [redacted]-aiquickactions- — Standard

Click Upload.	
Upload the ***ai-quick-action-finetuning-dataset.jsonl*** Which should come with this guide. The dataset must be in ***jsonl*** format.	 Choose Files from your Computer Drop files ai-quick-action-finetuning-dataset.jsonl 42.79 KiB
Now navigate to your Data Science Projects. ***OCI Menu > Analytics & AI > Data Science.***	 ORACLE Cloud Search resources, services, documentation, and Marketplace Search Home Compute Storage Networking Oracle Database Databases Analytics & AI Developer Services Identity & Security Observability & Management Hybrid Analytics & AI Analytics Analytics Cloud Data Intelligence Data Lake Big Data Service Data Catalog Data Integration Data Flow Messaging Streaming Connector Hub Machine Learning Data Science Data Labeling Data Science
Open up your existing Data Science Project.	 ORACLE Cloud Search resources, services, documentation, and Marketplace Data Science Projects Private endpoints List scope Compartment [redacted] [redacted] Projects *in* [redacted] *compartment* Data Science Prerequisites Show more information Create project Name State Description ismailsyed-dsp-1 ● Active -

Click on your existing Data Science Notebook Session. Note – This does not have to be a GPU Shape.	 Notebook sessions *in* [redacted] *compartment* Create notebook session **Name** **State** **Compute instance shape** **Networking** **Endpoint** ismailsyed-nbs-2 ● Inactive VM.GPU.A10.1 Custom Public ismailsyed-nbs-1 ● Active VM.Standard.E4.Flex Custom Public
Click on Open. This will open up your Data Science Notebook Session. ***You will have to reauthenticate.***	 Data Science » Projects » Project detail: Notebook sessions » Notebook session d ismailsyed-nbs-1 ACTIVE Notebook session informatio General information **OCID:** [redacted] Show Copy **Created on:** Wed, May 8, 2024, 1
If the policies within the pre-requisites have been implemented correctly you should be able to open up the AI Quick Actions Extension within the Launcher. ***Click AI quick actions.***	 Welcome to the Data Science service The Launcher provides easy access to your notebooks, console, text editor, termina To get started, use the Environment Explorer to install a conda environment. To be able to publish your own conda environments, specify the location to stor **AI quick actions** Test, deploy and fine-tune foundation models with AI quick actions **Environment Explorer** Explore and manage conda environments. 1-Demos 3 months ago 2-CustomerWork 23 days ago 3-InternalWork 20 days ago 4-Training 3 months ago 5-Resources 3 months ago conda 3 months ago

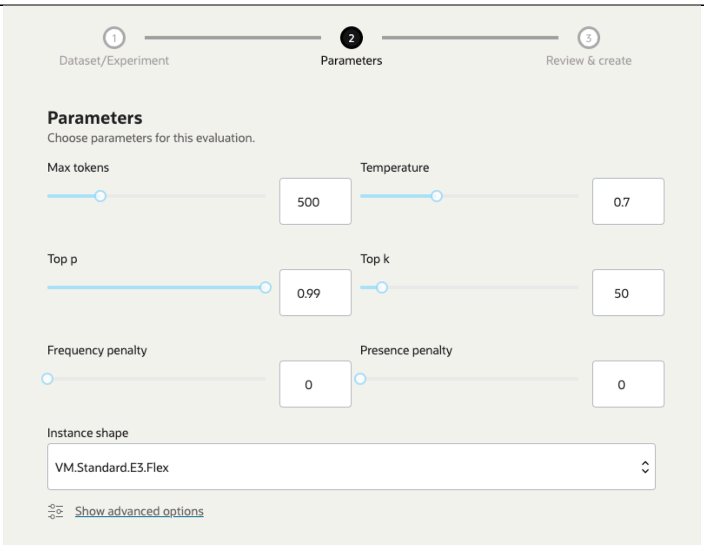
Click on the Evaluations Tab.	 Launcher AI quick actions AI quick actions Explore, fine tune, deploy, test and evaluate popular Large Language Models Models Deployments Evaluations Model explorer
Click on Create Evaluation.	 Evaluations Evaluations help you refine your models based on the metric output. For more details, visit our documentation site. Create evaluation Search evaluations Lifecycle state: Any state Display name · Evaluation source · Experiment · Lifecycle state · Created on No items found. Showing 0 items 1 of 1 AI quick actions
Select your Compartment. ***Enter a name for your evaluation.*** ***Select your existing deployed model.*** ***Enter an Evaluation Description.***	 1 Dataset/Experiment 2 Parameters 3 Review & create Model Deployment information Choose a deployment and add an optional description for this evaluation. Compartment: [redacted] Evaluation name: evaluation-isyed-mistral-7b-instruct-20240814 Model deployment name: isyed-mistral-7b-instruct-20240814 Evaluation description *Optional*: Evaluate Deployed Mistral 7b Instruct v0.1 LLM
Select your Evaluation Metric. ***I have selected Bert Score.*** You can select multiple evaluation metrics.	 Evaluation metric Choose evaluation metric to be computed as part of this model evaluation. metrics from the list of metrics supported. Evaluation metric: bertscore ×

Select your Evaluation dataset. ***Choose an existing dataset.*** ***Select your compartment.*** ***Select your Object Storage Bucket (Data Bucket)*** ***Enter path/name of your evaluation dataset.***	 ● Choose an existing dataset ○ Upload dataset from notebook storage Select compartment [redacted] Object Storage location ismailsyed-bucket-aiquickactions-data Object Storage path [redacted] ai-quick-action-finetuning-dataset.jsonl File extension must be .jsonl and size less than 100Mb
Define new Experiment. ***Select Create new experiment.*** ***Enter Experiment Name.*** ***Enter Experiment Description.***	 **Experiment** Choose an experiment for this evaluation. Or create a new experiment. ○ Choose an existing experiment ● Create a new experiment Experiment name experiment-mistral-7b-instruct Experiment description *Optional* Experiment for Mistral 7b Instruct
Define where to save evaluation results. ***Select Compartment.*** ***Select Object Storage Location (Model Bucket).*** ***Enter path where to save results.*** ***Click Next.***	 **Results** Choose bucket information to save the evaluation result. Select compartment [redacted] Object Storage location [redacted]-bucket-aiquickactions-models Object Storage path *Optional* [redacted] eval-mistral7b/ Must be a directory Close Next

Here you can define your parameters for the LLM. ***I have left as the default.***

Select your Instance shape to run the evaluation. ***I have left as the default. VM.Standard.E3.Flex.***

Click Next.



Parameters

Choose parameters for this evaluation.

Max tokens: 500

Temperature: 0.7

Top p: 0.99

Top k: 50

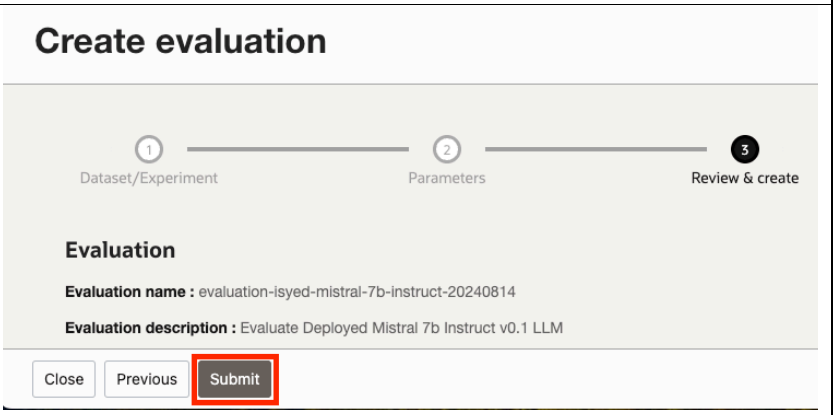
Frequency penalty: 0

Presence penalty: 0

Instance shape: VM.Standard.E3.Flex

Review all the Evaluation configuration.

Click Submit.



Create evaluation

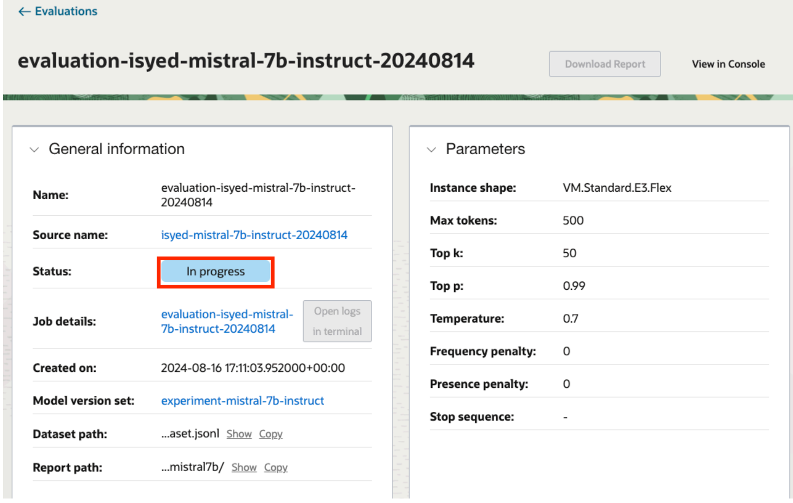
Evaluation

Evaluation name : evaluation-isyed-mistral-7b-instruct-20240814

Evaluation description : Evaluate Deployed Mistral 7b Instruct v0.1 LLM

This will kick start an OCI Data Science Job in the background to run the Evaluation.

The Evaluation Status will be displayed as ***In Progress.***



evaluation-isyed-mistral-7b-instruct-20240814

General information

Name: evaluation-isyed-mistral-7b-instruct-20240814

Source name: isyed-mistral-7b-instruct-20240814

Status: In progress

Job details: evaluation-isyed-mistral-7b-instruct-20240814

Created on: 2024-08-16 17:11:03.952000+00:00

Model version set: experiment-mistral-7b-instruct

Dataset path: ...aset.jsonl Show Copy

Report path: ...mistral7b/ Show Copy

Parameters

Instance shape: VM.Standard.E3.Flex

Max tokens: 500

Top k: 50

Top p: 0.99

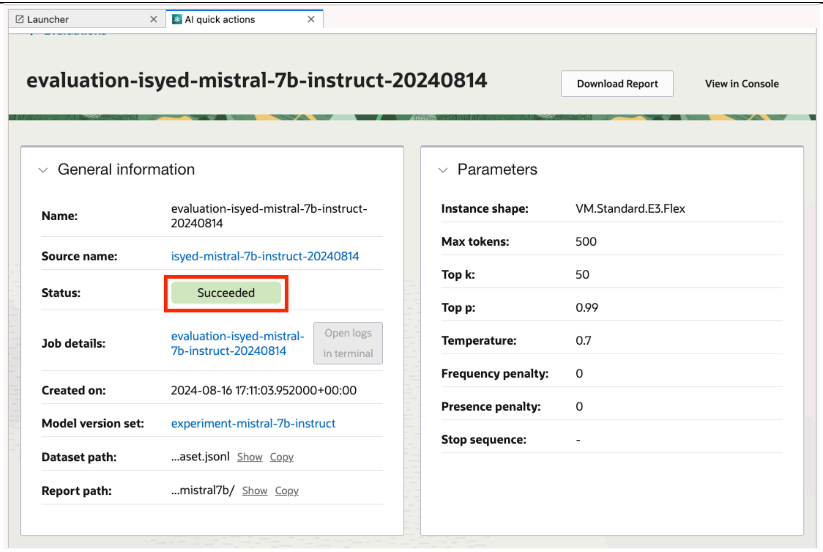
Temperature: 0.7

Frequency penalty: 0

Presence penalty: 0

Stop sequence: -

Once the Evaluation Job is completed the status will update to ***Succeeded***.



evaluation-isyed-mistral-7b-instruct-20240814

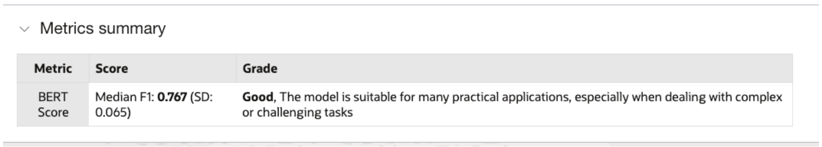
General information

Name:	evaluation-isyed-mistral-7b-instruct-20240814
Source name:	isyed-mistral-7b-instruct-20240814
Status:	Succeeded
Job details:	evaluation-isyed-mistral-7b-instruct-20240814
Created on:	2024-08-16 17:11:03.952000+00:00
Model version set:	experiment-mistral-7b-instruct
Dataset path:	...aset.jsonl
Report path:	...mistral7b/

Parameters

Instance shape:	VM.Standard.E3.Flex
Max tokens:	500
Top k:	50
Top p:	0.99
Temperature:	0.7
Frequency penalty:	0
Presence penalty:	0
Stop sequence:	-

If you scroll down, you will see the evaluation metrics displayed. In our case, the BERT Score.

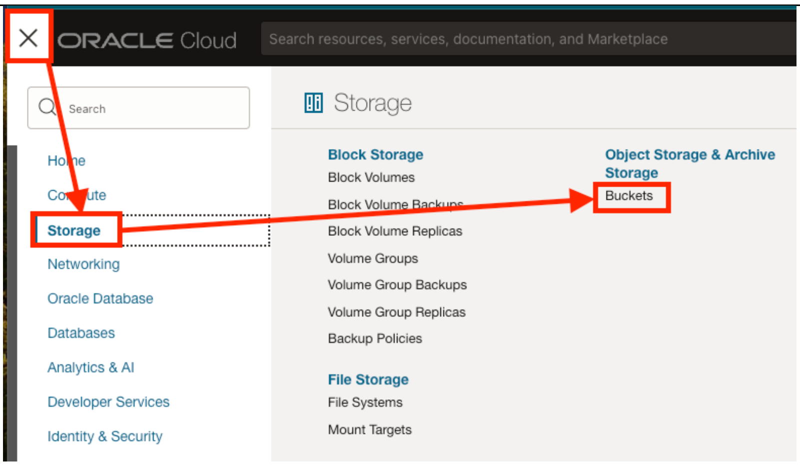


Metrics summary

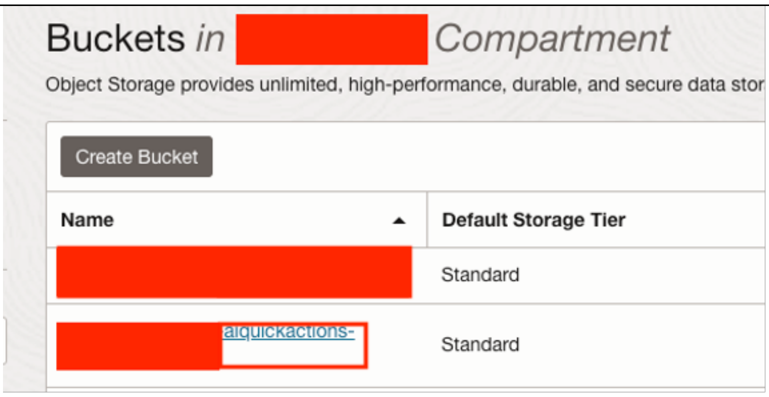
Metric	Score	Grade
BERT Score	Median F1: **0.767** (SD: 0.065)	**Good**, The model is suitable for many practical applications, especially when dealing with complex or challenging tasks

Head back to the OCI Console and visit your Object Storage Buckets.

Navigate to OCI Menu > Storage > Buckets.



Select your Models Object Storage Bucket which we used to save the Evaluation Metrics to.

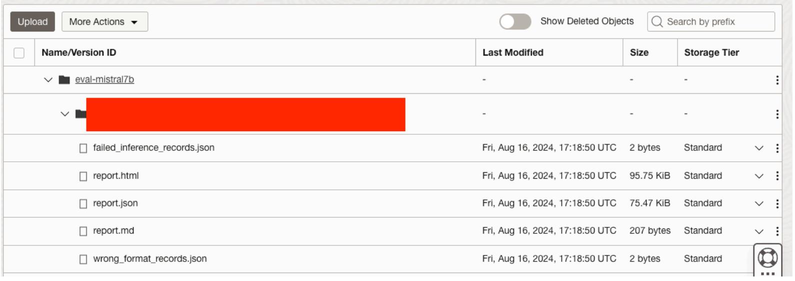


Buckets *in* [redacted] *Compartment*

Object Storage provides unlimited, high-performance, durable, and secure data stor

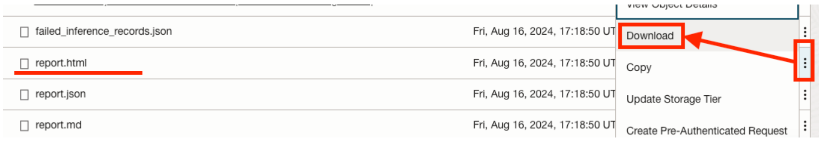
Name	Default Storage Tier
[redacted]	Standard
[redacted]	Standard

Under our Evaluation path, ***expand the results for our Evaluation Job Run.***



☐ Name/Version ID	Last Modified	Size	Storage Tier
eval-mistral7b	-	-	-
[redacted]	-	-	-
☐ failed_inference_records.json	Fri, Aug 16, 2024, 17:18:50 UTC	2 bytes	Standard
☐ report.html	Fri, Aug 16, 2024, 17:18:50 UTC	95.75 KiB	Standard
☐ report.json	Fri, Aug 16, 2024, 17:18:50 UTC	75.47 KiB	Standard
☐ report.md	Fri, Aug 16, 2024, 17:18:50 UTC	207 bytes	Standard
☐ wrong_format_records.json	Fri, Aug 16, 2024, 17:18:50 UTC	2 bytes	Standard

Use the 3 dots to Download the report.html file which gets automatically generated by the Evaluation Job Run.



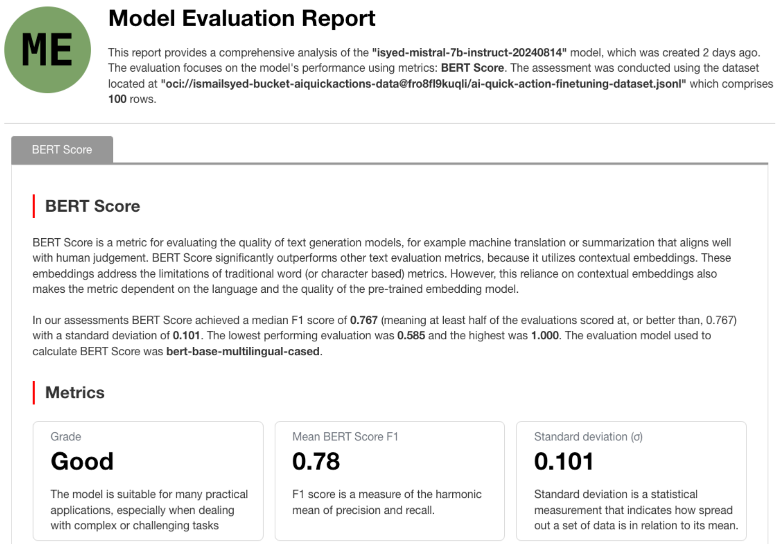
☐ failed_inference_records.json	Fri, Aug 16, 2024, 17:18:50 UT
☐ report.html	Fri, Aug 16, 2024, 17:18:50 UT
☐ report.json	Fri, Aug 16, 2024, 17:18:50 UT
☐ report.md	Fri, Aug 16, 2024, 17:18:50 UT

We can then take a look at the Model Evaluation Report.

Here we can see:

Description of the Evaluation Metric.

Overview of the Evaluation Metrics Calculated.



Model Evaluation Report

This report provides a comprehensive analysis of the **"isyed-mistral-7b-instruct-20240814"** model, which was created 2 days ago. The evaluation focuses on the model's performance using metrics: **BERT Score**. The assessment was conducted using the dataset located at **"oci://ismailsyed-bucket-aiquickactions-data@fro8fl9kuqli/ai-quick-action-finetuning-dataset.jsonl"** which comprises **100** rows.

BERT Score

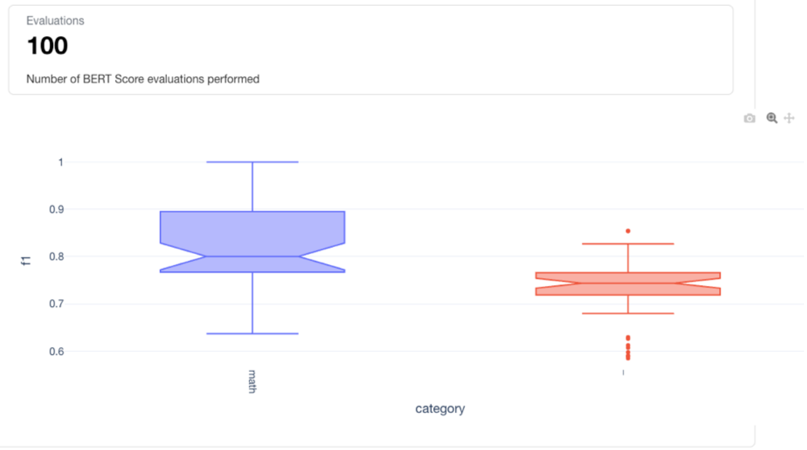
BERT Score is a metric for evaluating the quality of text generation models, for example machine translation or summarization that aligns well with human judgement. BERT Score significantly outperforms other text evaluation metrics, because it utilizes contextual embeddings. These embeddings address the limitations of traditional word (or character based) metrics. However, this reliance on contextual embeddings also makes the metric dependent on the language and the quality of the pre-trained embedding model.

In our assessments BERT Score achieved a median F1 score of **0.767** (meaning at least half of the evaluations scored at, or better than, 0.767) with a standard deviation of **0.101**. The lowest performing evaluation was **0.585** and the highest was **1.000**. The evaluation model used to calculate BERT Score was **bert-base-multilingual-cased**.

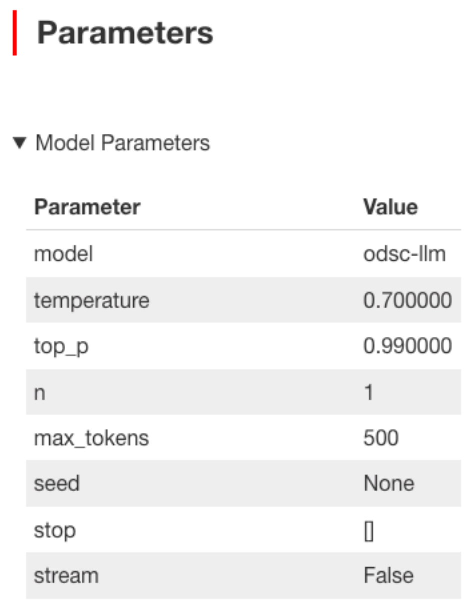
Metrics

Grade	Mean BERT Score F1	Standard deviation (σ)
Good	0.78	0.101
The model is suitable for many practical applications, especially when dealing with complex or challenging tasks	F1 score is a measure of the harmonic mean of precision and recall.	Standard deviation is a statistical measurement that indicates how spread out a set of data is in relation to its mean.

We also get a ***Box Plot of the BERT F1 Score broken down by the different Categories defined in our Evaluation Dataset.*** We can see it performed better on Math related questions compared to the NULL Category.



We get a list of all the parameters the Model was invoked with.

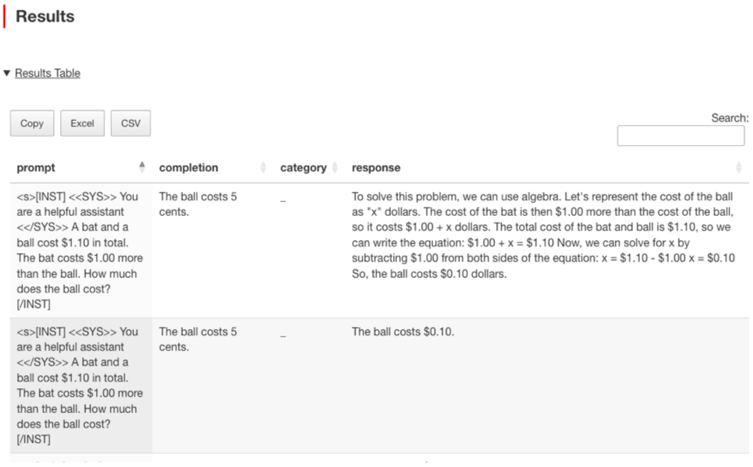


Parameters

▼ Model Parameters

Parameter	Value
model	odsc-llm
temperature	0.700000
top_p	0.990000
n	1
max_tokens	500
seed	None
stop	[]
stream	False

Finally, we get a list of each individual sample in our Evaluation Dataset with the Prompt vs the Completion, vs the Response generated by the Model.



Results

▼ Results Table

prompt	completion	category	response
<s>[INST] <<SYS>> You are a helpful assistant <</SYS>> A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? [/INST]	The ball costs 5 cents.	_	To solve this problem, we can use algebra. Let's represent the cost of the ball as "x" dollars. The cost of the bat is then $1.00 more than the cost of the ball, so it costs $1.00 + x dollars. The total cost of the bat and ball is $1.10, so we can write the equation: $1.00 + x = $1.10 Now, we can solve for x by subtracting $1.00 from both sides of the equation: x = $1.10 - $1.00 x = $0.10 So, the ball costs $0.10 dollars.
<s>[INST] <<SYS>> You are a helpful assistant <</SYS>> A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? [/INST]	The ball costs 5 cents.	_	The ball costs $0.10.