

## 1 INSTALLATION

### Pre-requisites

- ☐ Python 3.11
  - ☐ Node.js v20.11.1 LTS and above
  - ☐ Git
- We recommend using a Virtual Env.*

### Installation

```
pip install "aiverify-moonshot[all]"
python -m moonshot -i moonshot-data -i moonshot-ui
```

### Run Moonshot

```
python -m moonshot web
```

Open <http://localhost:3000/> in a browser.

## 2 CHOOSING TESTS

- Click [Get Started →](#)
- Select the cookbooks relevant to your use case.
- Click on [these cookbooks](#) to see test details and decide if you want to add/ remove any tests.
- Once done, proceed by clicking [✓](#)

## 3 CONNECTING AI SYSTEMS

To test models with existing endpoints in Moonshot,

- Click on [✎ Edit](#) for the model you wish to test.
- Provide your API Token and click [Save](#)

To test models with no existing endpoints in Moonshot,

- Click on [+ Create New Endpoint](#)
- Provide the following info and click [Save](#)

**Name** - A unique identifier for this new endpoint (Required)  
**Connection Type** - Type of model connector API to use (Required)  
**URI** - URI to the endpoint.  
**Token** - Your private API token.  
**Max Calls Per Second** - The max. no. of calls to be made per second.  
**Max Concurrency** - The max. no. of calls to be made at any one time.  
**Other Parameters** - Certain connector types require other parameters.  
 Tip: For OpenAI and Claude, you will need to specify the 'model'.

- Select the endpoints that you wish to test, and click [✓](#)



If you wish to run these cookbooks:

- MLCommons AI Safety Benchmarks v0.5
- Facts about Singapore

you'll need to add your API key for **Together Llama Guard 7B Assistant**.

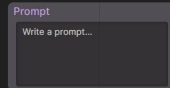
## 4 RUNNING BENCHMARKS

- Provide the following info and click [Run](#)
  - Name** - A unique identifier for this benchmark run.
  - Description** - Describe the purpose and scope of this run.
  - Run a smaller set** - The number of prompts per recipe to be run (Indicating 0 will run the full set)
- To view the progress of your run, click [🔔](#)
- Once the run is completed, click [View Report](#)
- Click on [model-1](#) to toggle the report displayed.
- You can also [Download HTML Report](#) and [Download Detailed Scoring JSON](#)

## 5 RED TEAMING

- Start from [👤 Start New Session →](#) or [Discover new vulnerabilities](#) [Start Red Teaming →](#)
- Select the endpoints that you wish to Red-Team.
- Select an attack module to try out and click [✓](#) or click [Skip for now](#)
- Provide the following info and click [Start](#)
  - Name** - A unique identifier for this red-teaming session.
  - Description** - Describe the purpose and scope of this session.

### Sending Prompts

Type your prompt in  and click [Send](#)

### Saving & Ending Sessions

All sessions are being saved in real time. You can click [✕](#) to exit a session any time.

### Red-Teaming tools available



#### Attack Modules

Techniques that enable the automatic generation of adversarial prompts.



#### Prompt Templates

Text structures that guide the formatting and contextualisation of the prompt sent.



#### Context Strategies

Approaches to append the session's context to the next prompt sent.

