AI VERIFY MOONSHOT
beta release

1 INSTALLATION

Pre-requisites

- ☐ Python 3.11
- ☐ Git
- ☐ Node.js v20.11.1 LTS and above
  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

1. Click `Get Started →`
2. Select the cookbooks relevant to your use case.
3. Click on **these cookbooks** to see test details and decide if you want to add/ remove any tests.
4. Once done, proceed by clicking ⌄

3 CONNECTING AI SYSTEMS

To test models with existing endpoints in Moonshot,

1. Click on `✎ Edit` for the model you wish to test.
2. Provide your API Token and click `Save`

To test models with no existing endpoints in Moonshot,

1. Click on `+ Create New Endpoint`
2. 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'.

3. 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**.
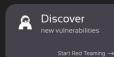
4 RUNNING BENCHMARKS

1. 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)

2. To view the progress of your run, click 🔔
3. Once the run is completed, click `View Report`
4. Click on `model-1 ⌄` to toggle the report displayed.
5. You can also `Download HTML Report` and `Download Detailed Scoring JSON`

5 RED TEAMING

Discover
new vulnerabilities

Start Red Teaming →

1. Start from `🤖 Start New Session →` or
2. Select the endpoints that you wish to Red-Team.
3. Select an attack module to try out and click ⌄ or click `Skip for now`
4. 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 [Prompt | Write a prompt...] 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.