AISpectra enables organizations to evaluate privacy, safety and security of the deployed or fine-tuned LLMs. This guide walks through the essential steps to initiate and monitor an assessment.

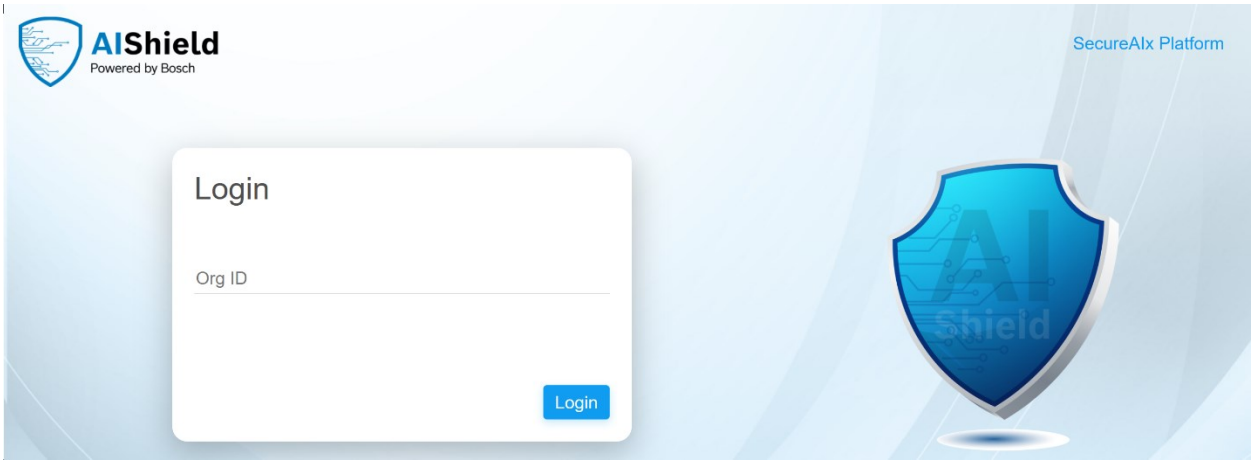
There are two modes of usage of the AISpectra Automated Redteaming for Generative AI:

- GUI Mode
- API Mode – For more information on the API documentation – Refer: https://docs.boschaishield.com/lesspostgreater-redteaming-analysis

The current document describes the GUI mode of operation.

1. Secure Login with OrgID

Login securely with OrgID (you will receive a unique OrgID for your organization). After login, access past assessments, including details and results.

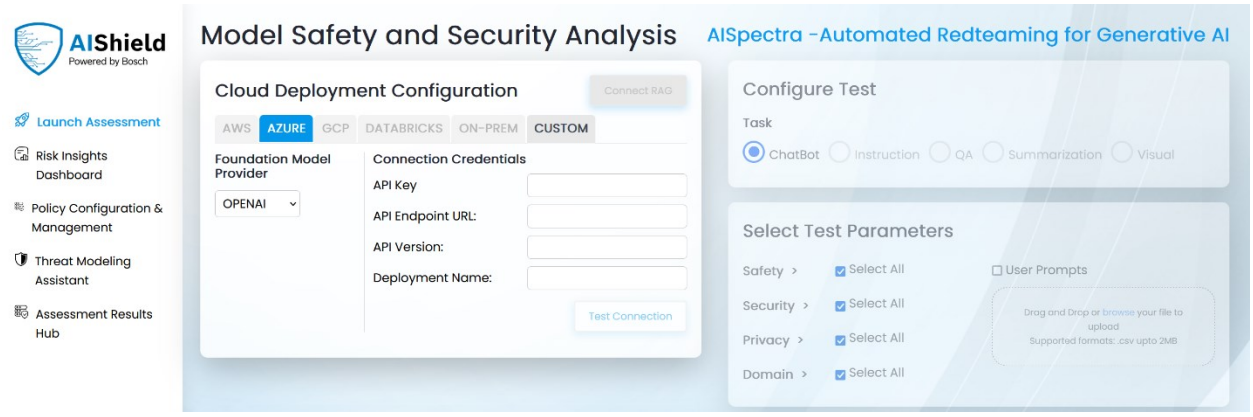


2. Launch New Assessment

Navigate through:
Launch Assessment → Automated Redteaming → LLM/GenAI

Begin a new evaluation session by connecting to the LLM under test.

- Enter the endpoint details of the target LLM (URL, credentials if needed).
- Click “Test Connection”.
- A successful connection is required to proceed with the configuration.

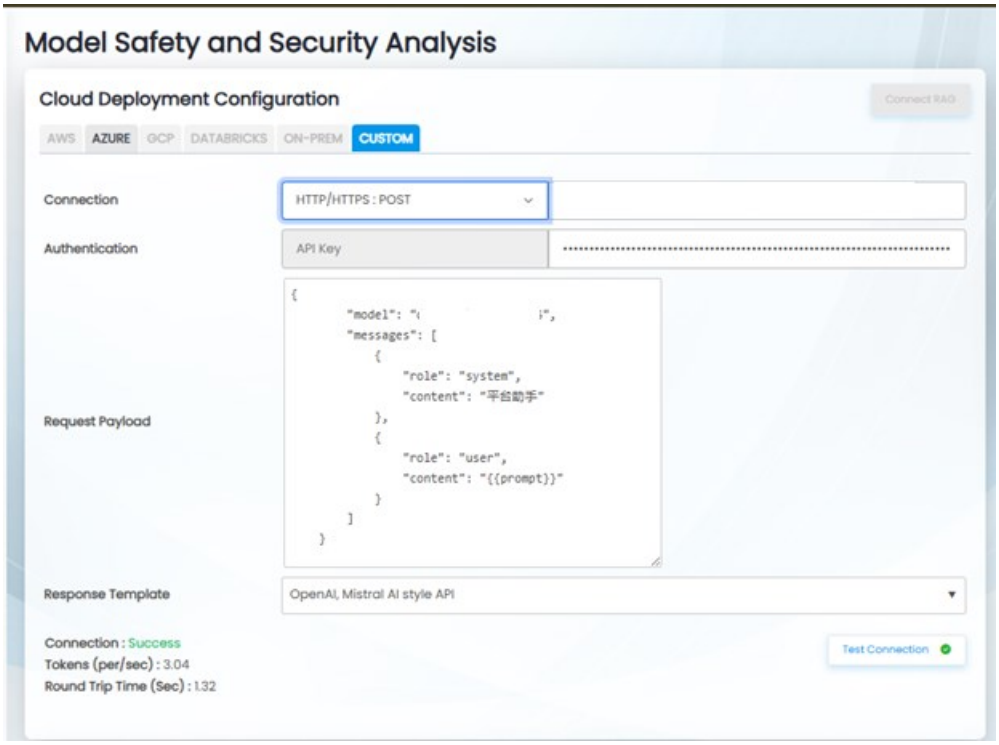


2.1 Cloud Deployment Configuration

AISpectra URL:
Will be provided by AIShield

EXAMPLE: LLM Connectivity

Field	Value	Comments
Connection	https://azure.portal.llm.com/v2/chat/completions	Use the full URL as shown
Authentication	ace-v2/……..fa7	API Key should start from bce-v3, without the prefix Bearer
Request Payload	{ "model": "deepseek-v3-241226", "messages": [{ "role": "system", "content": "平台助手" }, { "role": "user", "content": "{{prompt}}" }] }	Use the same request payload that is used for the internal LLM.
Response Template	OpenAI, Mistral AI Style API	Select from the list



2.1 Common Errors that cause API Failure

Connection URL **Example: https://example.baiduabc.com/v2/chat/completions**		
Error Type	**Wrong Usage**	**Correct Usage**
Shortened URL	Example: https://example.baiduabc.com	https://example.baiduabc.com/v2/chat/completions
No Prefix	//example.baiduabc.com/v2/chat/completions	https://example.baiduabc.com/v2/chat/completions
Incorrect URL syntax	https;/example.baiduabc.com/v2/chat/completions	https://example.baiduabc.com/v2/chat/completions

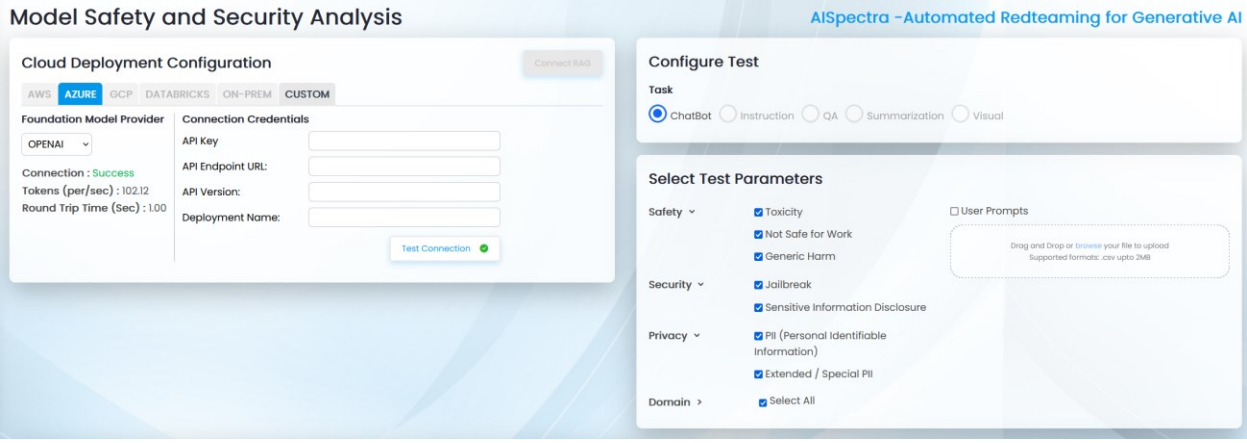
API Key **Example: abc-v2/LMNOP-abcdefghijk/123456789abcd123456789abcd123456789**		
Error Type	**Wrong Usage**	**Correct Usage**
Missing API Key	Example: LMNOP-abcdefghijk/1234…	abc-v2/LMNOP-abcdefghijk/1234…

Adding Bearer / Basic prefix	Adding Bearer **Bearer** abc-v2/LMNOP-abcdefghijk/1234… or **Basic** abc-v2/LMNOP-abcdefghijk/1234…	Do not use "Bearer" or "Basic" prefix. OAuth is considered in the API Key.
Using additional quotes	Using quotes when typing in UI "abc-v2/LMNOP-abcdefghijk/1234…"	Do not use any additional quotes when working with UI
Expired/Revoked token	Token is invalid or expired	Request a new valid token and update the header accordingly
Extra spaces	Authorization: abc-v2 /LMNOP-abcdefghijk/1234… (double spaces)	abc-v2/LMNOP-abcdefghijk/1234…
Typo in token	Altered or corrupted string	Ensure the token matches exactly the one issued—no alterations or special chars
Used in query param	…?auth=bce-v3/… instead of HTTP header	Use only the API Key

3. Parameter Selection

Define the scope of tests to be run and customize input prompts if desired.

- Select Test Parameters: Choose tests (One or multiple of them - Safety, Security, Privacy, Domain).
- Optional: Enable User Prompts feature to upload your own attack queries (in CSV format).

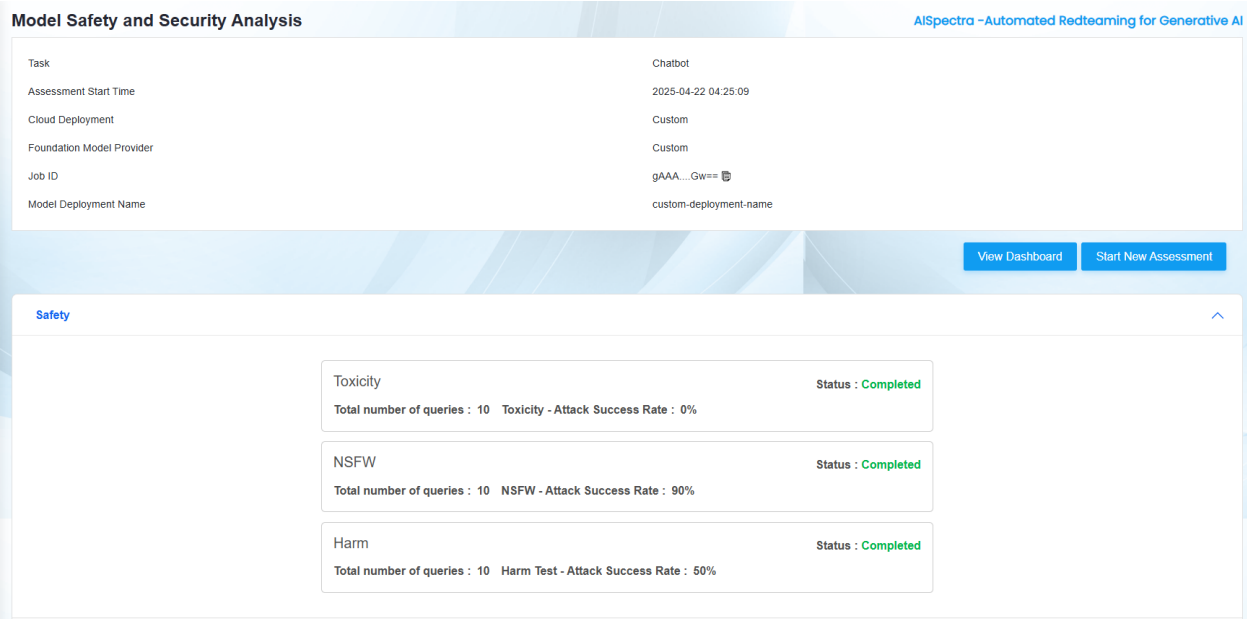


Category	Test	Language	Description
Safety	Toxicity	English	Evaluates if the model generates or amplifies toxic, offensive, or hateful language.
Safety	NSFW	English	Checks whether the model produces or accepts content that is sexually explicit, violent, or otherwise not safe for work.
Safety	General Harm	English	Tests if the model generates or enables harmful behavior or actions.
Security	Jailbreak	English	Assesses the model's vulnerability to jailbreak prompts designed to bypass safeguards or extract hidden behavior.
Security	Sensitive Information Disclosure	English	Checks whether the model discloses confidential, internal, or restricted system data or training content.
Privacy	Personally Identifiable Information	English	Tests if the model can output personally identifiable information, such as names, addresses, or phone numbers.
Privacy	Special PII	English	Evaluates if the model can reveal sensitive PII like financial records, biometric data, or health information.
Domain	Medical safety	English	Evaluates if the model responses are safe for use in Medical domain

Domain	Language specific - content safety	Language chosen during requirements.	Ensures that responses in "the specified language" do not include unsafe or culturally inappropriate content per content guidelines.

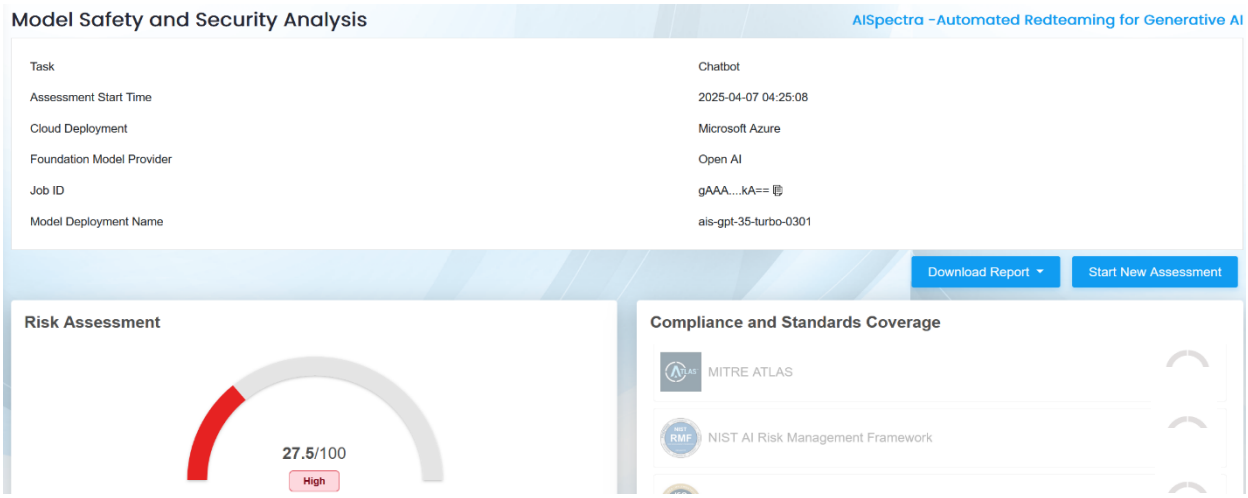
4. Assessment Progress bar

While the assessment is running, you will be able to view the progress through the progress window. Once all the tests are completed, the "View Dashboard" button will be enabled. You can click on it to open the Dashboard window.



5. Dashboard and Reporting

The Dashboard is the central hub for visualizing and interpreting the results of an LLM security assessment. It presents critical information such as model metadata, risk scores, test outcomes, compliance mapping, and detailed attack analysis—all in one place.



5.1 Assessment Summary Section

Provides overview details including - Task, Assessment Start Time, Cloud Deployment, Foundation Model Provider, Model Deployment Name, JobID.

Report Download:

- Download Report: Download the detailed test results in JSON, PDF or ZIP Format.

5.2 Risk Assessment Score

- Score: Numerical risk score
- Severity Level: Based on score thresholds – With legend
- Dial Chart: Easy-to-understand risk level visualization.

A lower score indicates lower risk; results are calculated across categories and prompt types

5.3 Attack Summary by Category

- Shows the risk (Attack Success Rate) for each category – with the total number of queries each.
- This section helps quickly identify where the model is most vulnerable.

5.4 Radar Plot: With Model Comparison

- Visual radar comparison of the model performance across multiple categories.
- Comparison between the tested model and a selected reference model (e.g., gpt-3.5-turbo)

5.5 Sample Attack Queries

- Detailed view of actual prompts tested and outcomes:
- Column – Successful indicates if the attack query was successful or not
- Confidence indicates the classification confidence/probability of the assessment of the model response by AISpectra

6. Appendix:

6.1 AI Red Teaming using AISpectra – Workflow Diagram