1. LLM / AI-specific Evaluation & Testing Frameworks

DeepEval is a simple-to-use, open-source LLM evaluation framework.

https://github.com/confident-ai/deepeval

promptfoo is a developer-friendly local tool for testing LLM applications.

https://github.com/promptfoo/promptfoo

Ragas is your ultimate toolkit for evaluating and optimizing Large Language Model (LLM) applications.

https://github.com/vibrantlabsai/ragas

LangSmith - LLM Evaluation

https://docs.langchain.com/langsmith/evaluation

trulens - LLM Evaluation framework

https://github.com/truera/trulens/

Phoenix is an open-source AI observability platform designed for experimentation, evaluation, and troubleshooting. (Tracing, Evaluation, Hallucination detection)

https://github.com/Arize-ai/phoenix

Langfuse is an open source LLM engineering platform.

https://github.com/langfuse/langfuse

Opik (built by Comet) is an open-source platform designed to streamline the entire lifecycle of LLM applications.

https://github.com/comet-ml/opik

LLM Canary tool is an easy-to-use open-source security benchmarking test suite.

https://github.com/LLM-Canary/LLM-Canary

Curated list of awesome open-source tools, resources, and tutorials for MLSecOps (Machine Learning Security Operations).

https://awesomemlsecops.com/

2. Adversarial & Robustness Testing Libraries

Adversarial Robustness Toolbox (ART) is a Python library for Machine Learning Security.

https://github.com/Trusted-AI/adversarial-robustness-toolbox

TextAttack is a Python framework for adversarial attacks, data augmentation, and model training in NLP.

https://github.com/QData/TextAttack

Foolbox: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX

https://github.com/bethgelab/foolbox

OpenAttack is an open-source Python-based textual adversarial attack toolkit.

https://github.com/thunlp/OpenAttack

3. Systematic Testing & Error Analysis Tools

Azimuth, an open-source dataset and error analysis tool for text classification

https://github.com/ServiceNow/azimuth

CheckList - Template-based testing for behavioral NLP model evaluation (minimum functionality, invariance, etc.).

https://github.com/marcotcr/checklist

PiML - An integrated Python toolbox for interpretable machine learning

https://github.com/SelfExplainML/PiML-Toolbox

OpenXAI : Towards a Transparent Evaluation of Model Explanations

https://github.com/AI4LIFE-GROUP/OpenXAI

4. Traditional ML Evaluation & Debugging Foundations

Deepchecks is a holistic open-source solution for all of AI & ML validation

https://github.com/deepchecks/deepchecks

OpenML- A worldwide machine learning lab

https://www.openml.org/

5. Collections

This repository packages a few benchmarks and agents used by the Center for AI Standards and Innovation (**CAISI)** for cyber capability evaluations of AI systems.

https://github.com/usnistgov/caisi-cyber-evals

Generative AI Red-teaming & Assessment Kit (NVIDIA)

https://github.com/NVIDIA/garak

Vulnhuntr leverages the power of LLMs to automatically create and analyze entire code call chains starting from remote user input and ending at server output for detection of complex, multi-step, security-bypassing vulnerabilities

https://github.com/protectai/vulnhuntr

Curated list contains 920 open-source projects for ML in python

https://github.com/lukasmasuch/best-of-ml-python

Curated list of open-source projects across ML, Web-Development, Crypto & Blockchain, Embedded, Robot development etc.

https://github.com/best-of-lists/best-of?tab=readme-ov-file

Curated collection of LLM apps built with RAG, AI Agents, Multi-agent Teams, MCP, Voice Agents etc.

https://github.com/Shubhamsaboo/awesome-llm-apps?tab=readme-ov-file

Curated list contains 410 MCP (Model Context Protocol) servers

https://github.com/tolkonepiu/best-of-mcp-servers

Collection of GPTs created by open-source community

https://github.com/taranjeet/awesome-gpts