

OPEN SOURCE SUMMIT INDIA

THE LINUX FOUNDATION



Open By Design - Open Ecosystem, Governed AI, Trusted Outcomes

Geeta Gurnani
Field CTO, IBM Technology,
India & South Asia

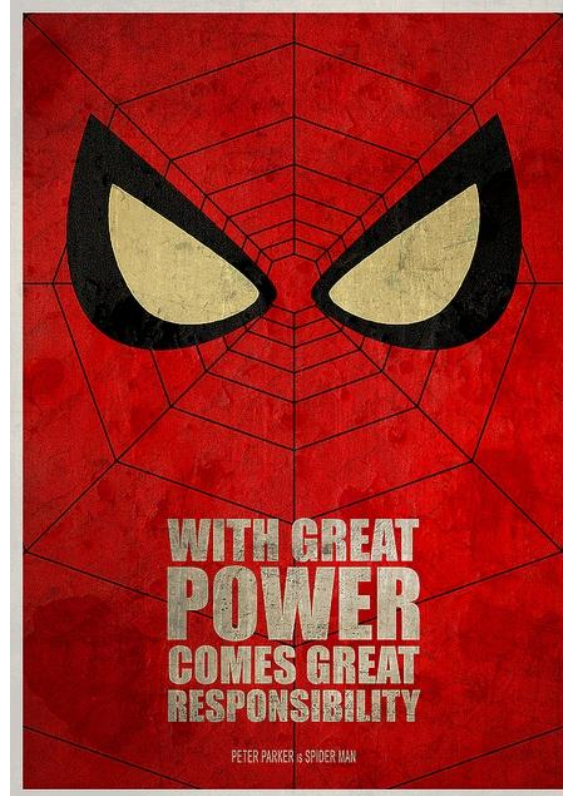


<https://www.linkedin.com/in/geetagurnani/>

#OSSUMMIT



Thank you for what you do to advance Tech



IBM and Open Source Communities

- Scaling Linux kernel for the enterprise
 - Very large cpu and memory configs, MM, FS,...
 - Observability: Tracing, profiling, Failure Logs,...
- Enterprise Hardening and security leadership
 - Common Criteria, OpenSSF, Open source supply chain



Project Lightwell



IBM and Open Source Communities

The Modern Era: AI, PyTorch, Data Preparation and Hardware Acceleration

LF AI & DATA



Docling

The mission of the project is to simplify document processing, to parse diverse formats — including advanced PDF understanding — and to provide seamless integrations with the gen AI ecosystem



BeeAI

The mission of the BeeAI project is to develop an open-source framework for building production-ready agents.

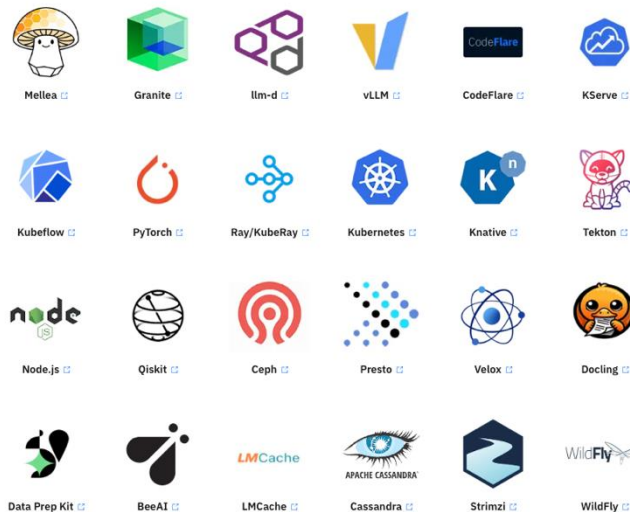


Data Prep Kit

The mission of the Data Prep Kit project is to simplify unstructured data preparation for LLM application development.

Build the future of tech with us

IBM is unmatched in the breadth of our open source involvement. From quantum and blockchain to containers, AI, and operating systems, we are actively leading in today's most influential projects and creating new projects to push technology forward for tomorrow. Join us in building the future with open source.



[Open Source @ IBM](#)

#OSSUMMIT

AI is writing code

AI is making decisions

AI is shaping business outcomes

Open Source is powering AI acceleration

But

Trust Gap is widening

Lack of explainability

Security vulnerabilities

Bias in decision-making

Hallucinations in generative models

Lack of explainability

Security vulnerabilities

Bias in decision-making

Hallucinations in generative models

Trusted AI = Responsible + Reliable + Transparent

Transparency

Fairness

Robustness

Accountability

Security & Privacy

Trusted AI = Responsible + Reliable + Transparent

Transparency

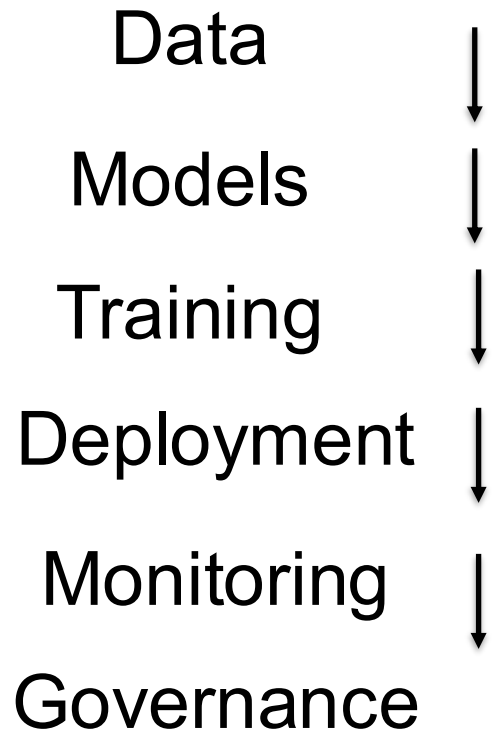
Fairness

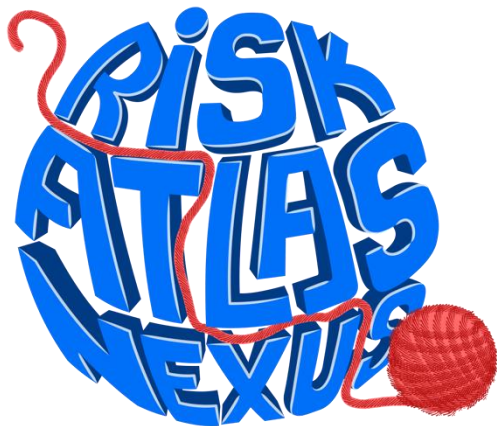
Robustness

Accountability

Security & Privacy

Embedded throughout the AI lifecycle





Traditional risk of AI Established risks of AI that apply to both traditional and generative models.

Specific to generative AI Risks that are specifically associated with generative AI models.

Amplified by generative AI Risks that are more severe or likely due to generative AI. These risks are also applicable to traditional AI models.

Specific to synthetic data Risks that are specifically associated with the use of synthetic data.

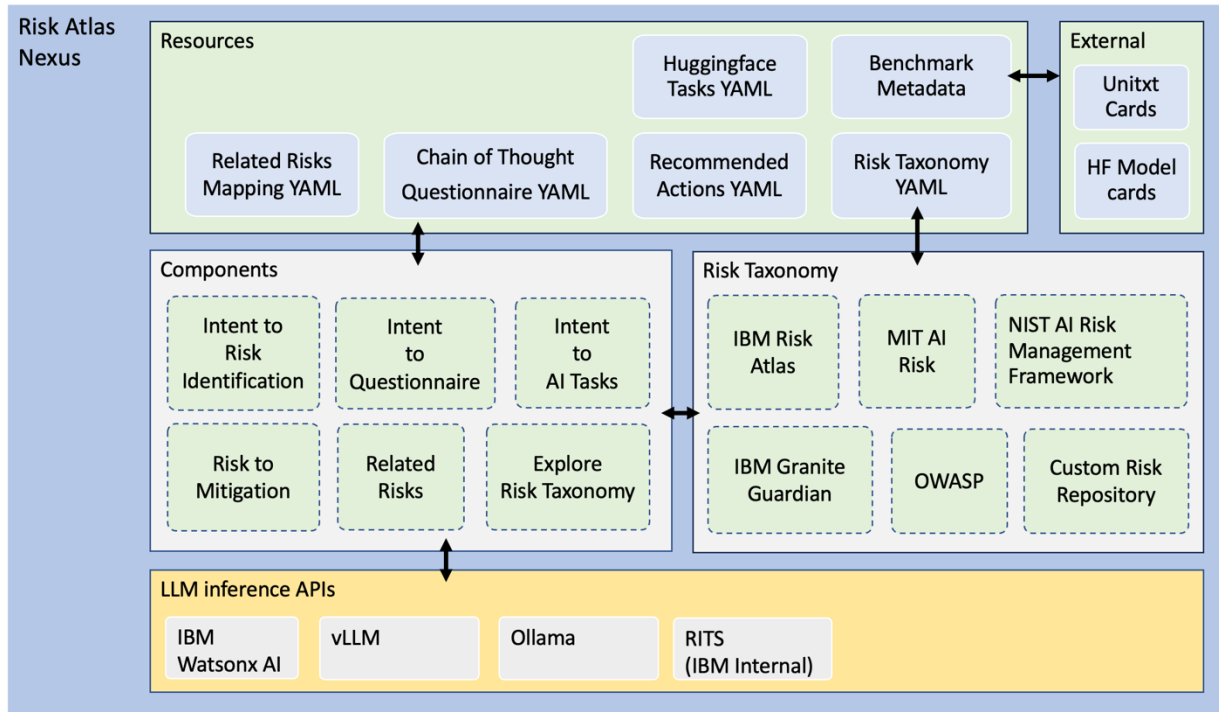
Amplified by synthetic data Risks that are increased or more likely to occur due to the use of synthetic data.

Navigate through the space of intrinsic based on your purpose

“Usage Governance Advisor: from Intent to AI Governance.” E. M. Daly, S. Rooney, S. Tirupathi, L. Garces-Erice, I. Vejbsbjerg, F. Bagehorn, D. Salwala, C. Giblin, M. L. Wolf-Bauwens, I. Giurciu, M. Hind, P. Urbanetz.
arXiv:2412.09157.



<https://github.com/IBM/risk-atlas-nexus>



Try it



<https://huggingface.co/spaces/ibm/risk-atlas-nexus>

“AI Risk Atlas: Taxonomy and Tooling for Navigating AI Risks and Resources.” F. Bagehorn, K. Brimijoin, E. M. Daly, J. He, M. Hind, L. Garces-Erice, C. Giblin, I. Giurgiu, J. Martino, R. Nair, D. Piorowski, A. Rawat, J. Richards, S. Rooney, D. Salwala, S. Tirupathi, P. Urbanetz, K. R. Varshney, I. Vejsbjerg, M. L. Wolf-Bauwens. arXiv:2503.05780.

Intent
Describe the intent of the application, or choose from one of the examples below.

I am creating a chatbot that will aim to provide factual information about health issues in Uganda.

Choose a risk taxonomy.
The risk taxonomy defines a wide range of risks, their classifications, and potential mitigations.

ibm-risk-atlas

Choose language model to use
Language model used to assess risks (This is not the model being assessed).

ibm/granite-20b-code-instruct

Example use cases

Medical chatbot Customer service agent

Submit

Potential Risks

Estimated by an LLM.

Harmful output Output bias Toxic output Jailbreaking Hallucination Evasion attack Incorrect risk testing

Over- or under-reliance Membership inference attack Confidential data in prompt Prompt leaking

Data privacy rights alignment Discriminatory actions IP information in prompt Legal accountability

Social hacking attack Indirect instructions attack Mitigation and maintenance AI agent compliance

Function calling hallucination Confidential information in data

Description:
A jailbreaking attack attempts to break through the guardrails established in the model to perform restricted actions.

Related Risks

Select a potential risk above to check for related risks.

Risks from other taxonomies related to atlas-jailbreaking

Jailbreaking Information Integrity LLM01:2025 Prompt Injection

Mitigations

Select a potential risk to determine possible mitigations.

Mitigation actions and controls related to risk atlas-jailbreaking.

Search...

Mitigation	Description
GV-1.2-001	Establish transparency policies and processes for documenting the origin and history of training data and generated data for GAI applications to advance digital content transparency, while balancing the proprietary nature of training approaches.

Open Source Community:

- Contribute to standards
- Build transparent AI systems
- Collaborate on (India) governance frameworks
 - M.** Moral and Ethical Systems
 - A.** Accountable Governance
 - N.** National Sovereignty
 - A.** Accessible and Inclusive
 - V.** Valid and Legitimate Systems

Open Source Community:

- Contribute to standards
- Build transparent AI systems
- Collaborate on (India) governance frameworks
 - M.** Moral and Ethical Systems
 - A.** Accountable Governance
 - N.** National Sovereignty
 - A.** Accessible and Inclusive
 - V.** Valid and Legitimate Systems

“Goldilocks” AI oversight

Towers
Different scopes of
alignment

Competence

Knowledge
Skills
Behaviors / Norms

(KSBs in military
parlance)

Transience

Semantic
(merged)

Episodic
(runtime)

Audience

Mass
Public
Small-group
Dyadic

“Scopes of Alignment.” K. R.
Varshney, Z. Ashktorab, D.
Bouneffouf, M. Riemer, J. D. Weisz.
Proceedings of the International
Workshop on AI Governance, Mar.
2025.



THE LINUX FOUNDATION
OPEN SOURCE SUMMIT
INDIA