

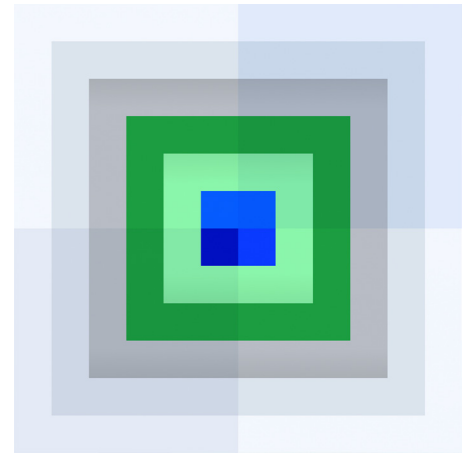
Responsible Use



Guide

Table of contents

1	Responsible Innovation
2	AI Ethics
3	How to Use This Guide
5	Building Responsible AI Systems
5	Cross-Functional Teams with Diverse, Technical Expertise Build Safer Models
5	AI Safety, Risk Mitigation and Management is at the Heart of Granite Development
6	AI Risk Atlas
6	Synthetic Data Generation for AI Safety
8	AI Safety and Alignment Workflow
13	Granite Guardian: The Next Step in Responsible AI
14	Sustainability and Energy Consumption
15	Appendix 1: IBM's Approach to AI Ethics
16	Appendix 2: Resources for Developers
18	Appendix 3: Usage Policies and Documentation
19	Appendix 4: Socio-Technical Harms and Risks



Responsible Innovation

Open innovation is the story of human progress. With a focus on building and opensourcing IBM Granite models, the IBM Granite team is leading the charge in building purpose-driven, cost-effective AI models for enterprise use. Comprised of AI engineers, research scientists, executives, ethicists, technical product managers, and other technologists and technical thought leaders, the IBM Granite team and its partners collectively form a responsible innovation ecosystem at IBM creating what's next in AI.

A commitment to responsible innovation practices underpins our entire approach. Most recently, the team has expanded our safety offerings with the release of Granite Guardian, a collection of models for safeguarding large language models (LLMs). This empowers model developers with a powerful, new, opensource AI safety tool. The foundations for the Granite Guardian models are instruction fine-tuned Granite language models of the same parameter size as the guardian. There is no other existing guardian model on the market today that comprehensively includes all risk dimensions covered by Granite Guardian.

IBM Granite models are used for a wide range of enterprise use cases. This includes text generation, classification, summarization, entity extraction, and other enterprise-specific applications. Accordingly, trust and safety must be built into these systems to apply them for enterprise use at scale. With ongoing research and testing internally, we expect that IBM Granite models will increasingly form the bedrock for agentic AI systems in future enterprise settings.

Whenever pushing the boundaries of what's next, it is important that responsible innovation drives strategic decision making. IBM's [Principles for Trust and Transparency](#), [Pillars of Trustworthy AI](#), [AI Ethics Governance Framework](#), and [AI Risk Atlas](#) are the cornerstones for our approach to safety and ethics. Internal data management practices, which allow for model and data lineage tracking, reinforce these best practices.

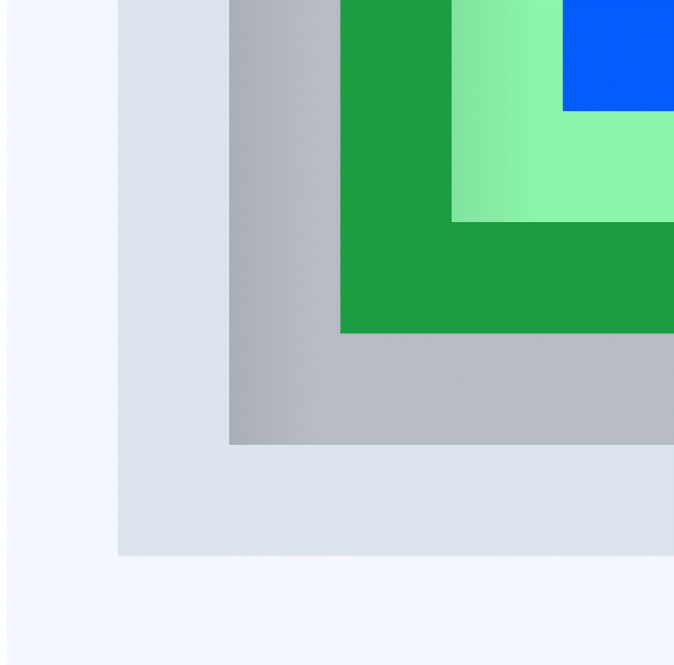
AI Ethics

Businesses are facing an increasingly complex, ever-changing global regulatory landscape when it comes to AI. The IBM approach to [AI ethics](#) aims to balance innovation with responsibility, helping businesses adopt trusted AI at scale.

The IBM AI Ethics Board is at the center of IBM's commitment to trust. The AI Ethics Board is a central, cross-disciplinary body that supports a centralized governance, review and decision-making process for IBM ethics policies, practices, communications, research, products and services. Its mission is to:

- 01** provide governance and decision-making as IBM develops, deploys, and uses AI and other technologies,
- 02** maintain consistency with the company's values, and
- 03** advance trustworthy AI for our clients, our partners, and the world.

The Board is a critical mechanism by which IBM holds our company and all IBMers accountable to our values and commitments to the ethical development and deployment of technology. For more information, refer to the AI Ethics Board's recent publication, [Foundation Models: Opportunities, Risks, and Mitigations](#). The paper [explores](#) the technology's benefits, risks, guardrails, and required risk mitigations.



Open innovation is the story
of human progress.

David Cox

Vice President of AI Models, IBM Research
IBM Director, MIT-IBM Watson AI Lab



How to use this guide

This guide is a resource for business executives, product managers, developers, and other AI practitioners seeking to leverage foundation models in a responsible way for enterprise use. The Guide covers contemporary AI safety choices faced by model developers, overviews risk mitigation tools and resources, provides energy calculations for the sustainable use of IBM's Granite models, and discusses other considerations for building and deploying responsible AI systems.

Tools, best practices, and recommended actions in this Guide are tailored for enterprise users of opensource foundation models. Nonetheless, we expect that the tools, best practices, and recommended actions herein will also inform broader corporate strategic decision making vis-à-vis the downstream deployment of AI technologies, capital allocation for AI safety projects, and ongoing discussion within public and private sectors on responsible AI systems.

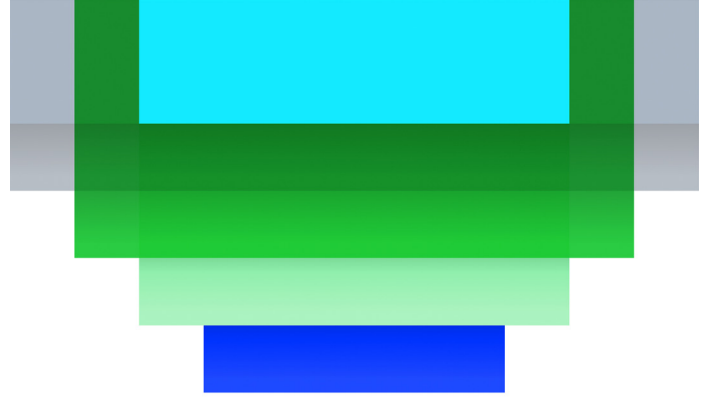
IBMer Spotlight

Kush Varshney

IBM Fellow

Kush R. Varshney trained as a signal processing and machine learning researcher. He has spent his entire 15-year professional career at IBM Research (primarily in Yorktown Heights, USA, but also in Nairobi, Kenya), starting as a postdoc and now as an IBM Fellow responsible for innovations in AI governance and watsonx.governance. Continuing his family history of using science and technology to advance social justice, he co-founded the IBM Science for Social Good initiative that paired IBM scientists, engineers, and student interns with social change organizations to help address poverty, hunger, and other inequalities of various kinds. He endeavors to carry fundamental basic research forward into practical impact, and make the tools and concepts developed along the way openly accessible. Toward this end, he led the creation of the AI Fairness 360 open-source toolkit and independently published the book “Trustworthy Machine Learning” available as a free pdf. Much of his current work is in collaboration with moral psychologists, lawyers, sociologists, and human-computer interaction scientists. At present, he serves as an IBM AI Ethics Board member, AAAI/ACM Conference on AI, Ethics and Society program committee co-chair, ACM Fairness, Accountability and Transparency Conference steering committee member, and IEEE Signal Processing Society distinguished industry speaker.





Building Responsible AI Systems

Cross-Functional Teams with Diverse, Technical Expertise Build Safer Models

Building responsible AI systems requires a holistic approach, bringing together cross-functional teams with diverse backgrounds. At IBM Research, this includes AI researchers and engineers, model developers, AI trust and safety technical experts, business executives, technical product managers, ethicists, legal and policy experts, data management experts, and model evaluation teams.

To implement its AI safety and governance work program, the IBM Granite team leverages one of its core, strategic assets: IBM's large, diverse, global, and technical employee base. This includes collaboration with the corporate InnerSource community to improve our instruction tuning and safety data and exposure to exploratory scientific research happening at IBM's global research hubs, like the [MIT-IBM Watson AI Lab](#).

Trust and governance must be at the core of enterprise AI.

Sriram Raghavan

Vice President, IBM Research AI

AI Safety, Risk Mitigation & Management is at the Heart of Granite Development

[AI risk management](#) is the process of systematically identifying, mitigating, and addressing the potential risks associated with AI technologies. It involves a combination of tools, practices and principles, with a particular emphasis on deploying formal AI risk management frameworks. AI risks can be model-specific, exist at system level, or be relevant to a given business use case.

Three key aspects underline this approach: AI ethics, trustworthy AI, and AI governance:

01 AI ethics refers to what an organization believes is the right thing to do, which is a societal decision, not a technological one.

02 Trustworthy AI involves technology that allows enterprises to measure models, detect bias, and make models more robust.

03 AI governance is about controls and processes, with the goal of ensuring you know which data went into your model, which tests were performed, and how best to monitor and iteratively improve model behavior post-deployment.

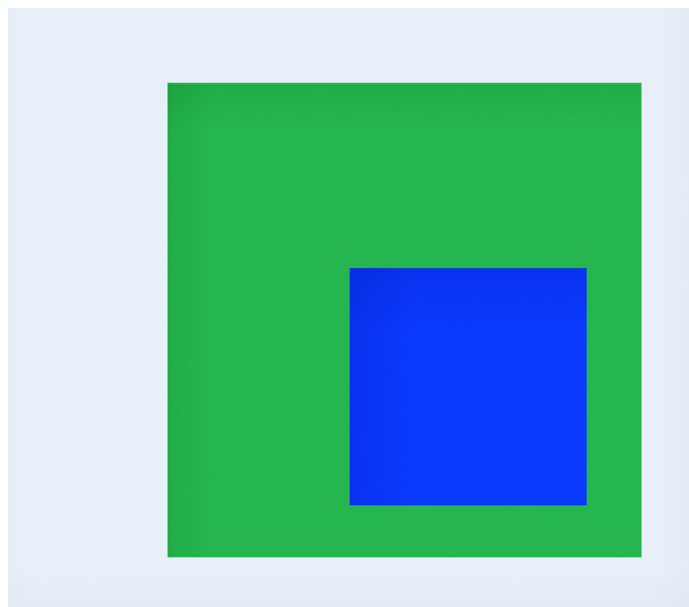
AI Risk Atlas

[The AI Risk Atlas](#) covers risks associated with inputs, outputs, and non-technical risks (e.g., governance, legal, and societal impact). It provides a guide to corporate decision makers looking to understand AI model risks. Risks are broadly categorized with one of three tags:

01 Traditional AI risks (applies to traditional models as well as generative AI)

02 Risks amplified by generative AI (might also apply to traditional models)

03 New and emerging risks (specifically associated with generative AI)



Synthetic Data Generation for AI Safety

Synthetic data can be a powerful source for augmenting AI safety training and alignment data in a targeted way. That is why the IBM Granite team leverages a centralized data generation and transformation framework to ensure a standardized process to take seeds (input-response pairs) and generate synthetic data. This synthetic data pipeline allows our research teams and subject matter experts with technical and domain expertise to create synthetic data for AI safety purposes.

An AI risk taxonomy is a repository that classifies and structures categories of risk. IBM Research uses its AI risk taxonomy as part of a broader set of safety measures that apply to its development of foundation models. The taxonomy helps to categorize known risks, which are used to generate synthetic data, for the purpose of aligning Research's foundation models.

Seeds are organized in a taxonomy that covers seven high-level categories:

01 Malicious Use: illegal activities, unethical or unsafe actions, and violence and extremism.

02 System Risks: security and operational risks.

03 Information Hazards: sensitive and personal information.

04 Discrimination: a wide range of discrimination, including implicit and explicit bias.

05 Societal Risks: disinformation, propaganda, and voter suppression.

06 Human-Chatbot Interactions: mental health, child harm, and self-harm.

07 Multi-Modal Requests: various forms of undesirable requests related to multi-modal support.

Lower-level nodes provide detailed coverage for each category. The taxonomy helps in carefully crafting and curating seeds containing safety data aligned with industry-specific, model-specific, and other types of risk. Our taxonomy has been informed by internal research as well as open-source AI risk taxonomies research conducted by the Massachusetts Institute of Technology (MIT)¹, MLCommons², and others³. Once generated, quality and consistency checks of the synthetic data are applied. This includes both extensive automated and manual review.

IBMer Spotlight

Nathalie Baracaldo

Manager of AI Security and Privacy Solutions
Senior Research Scientist, Master Inventor, Ph.D.

Nathalie Baracaldo is a Senior Research Scientist at IBM's Almaden Research Center in San Jose, CA, where she leads research projects related to AI safety. Nathalie is passionate about delivering machine learning solutions that are highly accurate, withstand adversarial attacks and protect data privacy. She currently focuses on securing LLMs using techniques such as unlearning and red teaming. Nathalie has a track record leading AI security and privacy projects, including being Principal Investigator for the DARPA program "Guaranteeing AI Robustness Against Deception" (GARD). Her team has performed red teaming and extended the Adversarial Robustness Toolbox (ART) that received the "Graduated Distinction" award by the Linux Foundation's AI & Data Foundation. Nathalie's research has received more than 5,000 citations on Google Scholar and multiple best paper awards. She also co-edited the book "Federated Learning: A Comprehensive Overview of Methods and Applications." Nathalie is a frequent keynote speaker and panelist. Nathalie twice received the IBM Master Inventor distinction for her contributions to IBM's intellectual property. Nathalie received her Ph.D. from the University of Pittsburgh in 2016.



AI Safety and Alignment Workflow

Key Terms

[Large language models](#) (LLMs) are a category of foundation models trained on immense amounts of data making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks. This is in stark contrast to the previous paradigm of building and training domain-specific models for each enterprise use case, individually. At high-level, the AI safety and alignment workflow generally maps to the model development lifecycle from data acquisition, storage, curation, and cleaning to training and model evaluation.

AI safety is the study and practice of how to deploy and operate AI in a manner that is beneficial to humanity and protects user safety at the individual (micro), societal (macro), and environment levels. Examples of core AI safety issues include: AI misuse, value misalignment, over-reliance on AI systems, hallucinations, and model bias.

AI alignment is the process of encoding human values and goals into large language models to make them as helpful, safe, and reliable as possible. Through alignment, enterprises can tailor AI models to follow their business rules and policies. More technical-oriented readers can find additional details on our training and alignment process in the *Granite 3.0 Language Models* white paper.

The IBM Granite models are produced by people, processes, and tools known collectively as the IBM Data and Model Factory.

Step 1: Data Preparation

Data Acquisition

Everything begins with data. IBM's data acquisition process uses a common, transparent, and controlled process to obtain data and make it available for training, tuning, alignment, and other downstream tasks in a standard format. This process is used for publicly available data as well as proprietary data obtained from data owners. It captures comprehensive information about a dataset, such as the data source and owner, how and when it was acquired, geographic location from where the data originated, licensing information (if applicable), and other meta-data. This process emphasizes quality, security, and ethical considerations. Furthermore, it is designed to avoid pirated materials, by excluding websites and datasets known to contain or disseminate such information. This process ensures that data is acquired in accordance with IBM's ethics principles.

In parallel to IBM's acquisition pipeline, IBM works with independent data owners, emphasizing quality, security, and ethical considerations.

Data Platform

IBM has robust processes and systems for data storage, data lineage tracking, and data management, post data acquisition. The platform is a comprehensive solution for managing datasets and models, offering complete visibility and analysis throughout the entire lifecycle—from data acquisition to utilization by various models, as well as governance and risk assessments.

Furthermore, it provides valuable insights into datasets and supports a robust model management framework, enabling effective and secure, access-controlled management of both models and datasets. The platform ensures that IBM foundation models work adheres to and ensures trust, governance, and openness for our enterprise clients. The platform is also designed and implemented to support substantive Governance, Risk and Compliance (GRC) activities that global enterprises face with AI.

The following functional and technological components are covered:

- a petabyte-scale Lakehouse with comprehensive user interface
- classification of massive data sources by jurisdictional and industry provenance
- required Python libraries, command-line interface (CLI), and relevant REST APIs
- large-scale processing capabilities for Spark and Ray clusters
- ad hoc data visualization & reporting capabilities for users

Broadly, it is the single source of truth and provides:

- a central repository for all datasets (public or private for training, fine tuning, alignment, instruction tuning, and evaluation for code and language models) and models
- insights into the datasets that enable accurate risk-based assessments
- intermediate results from processing the datasets
- lineage tracking of datasets and models
- model sharing mechanism between teams, including approval workflow
- model checkpoints and checkpoint evaluation results, incl. leaderboards
- a platform for FM-Eval based model evaluation for product infusion and other teams
- a platform for logging experiments and results and comparing and visualizing them (e.g., ablation studies, instruct tuning iterations etc.)
- a platform allowing users to create their reporting for leaderboards, experiments, or to meet regulatory compliance, etc.
- synthetic data store and lineage tracking along with model evaluation results and model management

Data Governance

In parallel, before releasing data for downstream use, a data clearance process assures that datasets are not used to train IBM foundation models, including the IBM Granite model series, without first taking careful consideration. Before data is added to IBM's curated pre-training dataset, it is subject to legal and governance review. The clearance request considers the acquisition metadata of a dataset as well as its planned purpose, usage restrictions, and sensitivity (e.g., personal information).

Once a dataset clearance review is complete, the dataset is tagged for potential inclusion in the training data mix, its metadata is moved into a catalog of approved datasets, and it is prepared for subsequent pre-processing stages.

Data Curation and Cleaning

IBM Research developed and open-sourced the [Data Prep Kit](#), a framework and toolkit for the preparation, annotation, and filtering of unstructured data (incl. language and code). The Data Prep Kit offers implementations of commonly needed data preparation steps, called transforms or modules, and a framework through which these transforms can be chained together to form data processing pipelines for end-to-end processing of unstructured data. Most recently, IBM Granite team leveraged the Data Prep Kit for preparing the large-scale training data for Granite 3.0 language models. Specifically, researchers used the Data Prep Kit to transparently scale data processing modules from a single laptop to a large cluster; provide lineage tracking of the processing jobs and logging of metadata to IBM Research's data platform (described above); and checkpoint capability for recovery from failures.

In addition, the IBM Granite team has capabilities that enable data processing and filtering for undesirable content. This includes hate, abuse, and profanity (HAP) content, content from pirated and other harmful sources, malware in code data, and many other forms of undesirable content across data modalities. This enables transparency around the data ingested into IBM Granite models.

Step 2: Training and Alignment

Pre-training teaches an LLM to continue generating text based on a given input. However, in practice, users often expect the LLM to treat the input as instructions to follow. To enable instruction following, we perform [supervised fine-tuning](#), with a mixture of datasets from different sources that have gone through the data clearance process. Each sample consists of a prompt and an answer with optional context. Supervised fine-tuning with IBM-generated synthetic data is designed to improve the model's conversational, safety, and instruction following capabilities.

We further train our SFT models with [reinforcement learning](#) for human preference alignment. This is a technique that aligns LLMs with human preferences, in which a trained reward model is used to optimize the performance of an LLM through reinforcement learning. This technique is uniquely suited for tasks with goals that are complex, ill-defined, or difficult to specify.

Importantly, the alignment data composition plays a critical role in the usefulness and behavior of language models. We primarily use publicly available high-quality datasets with permissive licenses, including synthetic prompts tailored for improving specific capabilities like knowledge-based question and answering. We perform several small-scale experiments to find the optimal mixture across four key categories, such as general English, code, math and safety.

Technical users should refer to the *Granite 3.0 Language Models* white paper for a technical deep dive.

Step 3: Model Evaluation and Vulnerabilities Screening

During training, model checkpoints are routinely tested against a suite of benchmarks for performance, robustness, and other metrics. Upon training completion, additional testing, including manual review by human evaluators prompting the model can help to identify potential weaknesses and vulnerabilities.

FMEval

The Foundation Model Evaluation Framework ([FMEval](#)) aims to validate and evaluate new LLMs coming out of the IBM Data and Model Factory in a systematic, reproducible, and consistent way. FM-eval supports both fine-tuning and prompting (in-context learning) evaluation modes, while providing out-of-the box academic as well as business benchmarks. FM-eval evaluates models in a modular way, starting during model training, with a basic evaluation, to get a quick indication of the model status, followed by a more comprehensive evaluation (i.e., more datasets, more templates, more seeds), and finally a complete evaluation (incl. harmfulness, robustness, privacy, and more).

unitxt

FM-eval is also designed to be flexible and allows easy addition of tasks, datasets, and metrics. To support this property, we developed [unitxt](#), an open-source Python library that provides a consistent interface and methodology for defining datasets, including the preprocessing required to convert raw datasets to the input required by LLMs, and the metrics used to evaluate the results.

The increasing versatility of LLMs has given rise to a new class of benchmarks that comprehensively assess a broad range of capabilities. Such benchmarks are associated with massive computational costs, extending to thousands of GPU hours per model. Thus, the evaluation team is looking into [efficient evaluation](#), where the goal is to intelligently reduce the computation requirements of LLMs evaluation, while maintaining an adequate level of reliability.

Internal and External Red Teaming

[Red teaming](#) is a process for testing model vulnerabilities, during which ethical hackers attempt to break the model – in effect, force the model to act in unexpected and potentially unethical ways. This may include generating undesirable content via adversarial and [prompt injection attacks](#).

IBM Research leverages both internal and external, automated and manual, red teaming techniques to screen for and address weaknesses. To ensure broad coverage of potential threat vectors, diverse communities are tapped for red teaming exercises.

In addition, we partner with a third-party company to do external red teaming of IBM Granite models, including our newly release IBM Granite 2B and 8B models. Algorithmic red teaming provides fast, actionable insights into potential model vulnerabilities from a trusted, external collaborator.

Internally, a permissioned playground environment allows select IBMers to probe Granite models and pre-release checkpoints for vulnerabilities and weaknesses. To gather insights, diverse communities within IBM with various domain expertise are granted access.

For example, an ongoing joint project between the Data and Model Factory at IBM Research and the Open Innovation Community (OIC)⁴ allows volunteer IBMers from around the globe to identify model weaknesses and contribute to the safety of our models.

Step 4: Continuous Feedback and Iterative Improvement

Finally, once built, it is important to systematically collect, triage, prioritize, and resolve feedback requests to continuously improve model performance. This can include requests for additional knowledge and skills. A robust feedback process supports model development, surfaces model weaknesses, and ensures that our research teams can make timely, iterative improvements to our models during training to improve safety performance.

⁴The OIC is an all-volunteer community at IBM where participants work on projects that benefit the company, our employees, our clients, and the world.

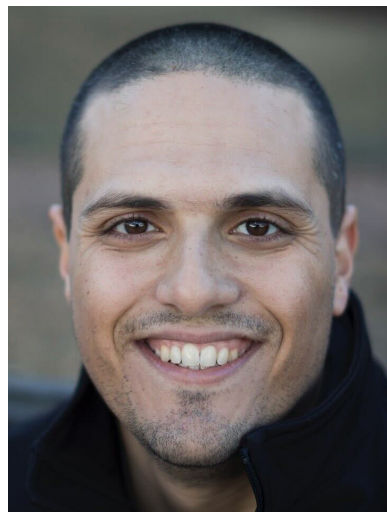
IBMer Spotlights



Ioana Baldini

Senior Research Scientist.

Ioana Baldini is a Senior Research Scientist in IBM Research AI. She is currently the tech lead for red teaming language models at IBM Research, supporting model evaluation and safety training data generation. As a result of collaborations with non-profit organizations, under the [IBM Science for Social Good Initiative](#), Ioana has focused her research efforts on natural language processing with the potential for social impact. Previously at IBM Research, Ioana was also part of the core research team that developed [OpenWhisk](#), an open source serverless platform, which is offered as [IBM Cloud Functions](#). In that role, Ioana contributed several components to the OpenWhisk infrastructure and helped productize it as the official IBM serverless platform. Ioana holds a Master's and PhD from the University of Toronto. She has received the NSERC (NSF-equivalent in Canada) Canada Graduate Scholarship, the IBM PhD Fellowship, the Canada Google Anita Borg Scholarship, and the IBM Research Division Award (for contributions to OpenWhisk).



George Kour

Research Scientist

George Kour is a researcher who specializes in measuring the behavior of AI models, with a focus on identifying and mitigating vulnerabilities related to the safety and security of LLMs. He feels passionately that technology holds incredible potential to address social inequalities, empower marginalized communities, and bridge divides between people, as witnessed first-hand, in 2016, during a volunteer opportunity to work with Syrian refugees in Greece. His work at IBM Research in the safety domain includes evaluating and unveiling vulnerabilities of LLMs and developing red teaming and guardrail techniques to enhance model security and reliability. He has led key initiatives, such as creating the AttaQ and ProvoQ safety benchmarks for internal and open-source evaluation frameworks. Kour's recent research explores the security of LLM-based agents, including attacks compromising their reasoning mechanisms. As we look to the future, we have an unprecedented opportunity and responsibility to shape the next, AI-driven technological era to uplift everyone, regardless of background or circumstance. George is excited to be part of a team at IBM Research creating that future.

Granite Guardian, the next step in responsible AI

Despite implementing robust, internal AI safety and governance measures, there will always be risk. Effective risk mitigation does not mean zero risk. Ideally, users should take a ‘Swiss cheese model’ of safety, adopting multiple layers of safeguards in their deployed AI systems, each with different ‘holes.’ Toward this end, the IBM Granite team provides a family of opensource Granite Guardian models that check input prompts and contexts, as well as language model outputs/responses for many of the harms catalogued in the IBM AI Risk Atlas.

At present, the Granite Guardian 2B and 8B variants check for social bias and implicit hate, toxicity and explicit hate, profanity, violence, sexual content, unethical behavior, jailbreaking, hallucination/groundedness, and answer relevance and context relevance (for retrieval-augmented generation use cases). An output of “yes” indicates a harmful response, and an output of “no” indicates that there is not harmful content in the response.

Additionally, the Granite Guardian models may be customized for other risk dimensions through prompting with the Build Your Own Detector functionality. The foundations for the Granite Guardian models are instruction fine-tuned Granite language models of the same parameter size as the guardian. They have been further trained on unique data, including human annotations

from socioeconomically diverse people and synthetic data generated using older Granite models, seeded by internal red teaming. There is no other existing guardian model on the market today that comprehensively includes all risk dimensions covered by Granite Guardian. For example, jailbreaking, hallucination, and RAG metrics are generally treated separately. This is important because it is understood within the research community that independent guardian models outperform general-purpose language models in harm detection.

Given their parameter size, the main Granite Guardian models are intended for use cases that require moderate cost, latency, and throughput, such as model risk assessments, model observability and monitoring, and spot-checking inputs and outputs. For use in data curation and real-time guardrailing with stricter cost, latency, or throughput requirements, the Granite Guardian collection also includes small detectors for single harm dimensions, such as a 38M-parameter model for recognizing hate, abuse and profanity. Model risk assessments may be facilitated by red teaming datasets opensourced by IBM Research, namely AttaQ, ProvoQ, and Social-StigmaQA.

Granite Guardian models may be used in conjunction with any open-weight or closed-weight language model, not only IBM Granite models.

Sustainability and Energy Consumption

At IBM, our [commitment](#) to sustainability dates to the 1970s, when we first established formal goals around energy conservation and waste management. Training foundation models consumes energy, resulting in emissions of carbon dioxide⁵. That is why IBM Granite models are trained on [Blue Vela](#), powered by 100% renewable energy.

Table 1: Estimated energy consumption and carbon emissions for training Granite 3.0 models

Model	GPU power consumption	GPU-hours	Total power consumption (MWh)	Carbon (tCO ₂ eq)
Granite 3.0 2B	700W	192,030	147.8	57.6
Granite 3.0 8B	700W	832,102	640.6	249.8
Granite 3.0 1B-400M	700W	71,171	54.6	21.3
Granite 3.0 3B-800M	700W	133,308	102.6	40.0

Looking ahead, we expect various mitigation strategies can be used to further reduce the energy consumed and carbon footprint of training future Granite models. For example, the amount of resources used in training may be adjusted as a function of the availability of renewable energy, or resource usage may be capped to not exceed certain energy usage or emissions limits.

Moreover, we hope that releasing all Granite 3.0 models in open source will help to reduce future carbon emission since training is already done. As the models are relatively lightweight, they can be run on a single GPU (maximum 8B parameters).

⁵To calculate Watt-hour, we use the formula: Wh = GPU-hours × (GPU power consumption) × PUE, where Power Usage Effectiveness (PUE) is set with 1.1. To calculate emissions, we use the US national average carbon intensity factor of 0.39 kg CO₂eq/KWh, according to U.S. Energy Information Administration, without taking location of data centers in consideration.

Appendix 1:

IBM's Approach to AI Ethics

IBM's multidisciplinary, multidimensional approach to AI ethics is built upon Principles for Trust and Transparency. The Principles for Trust and Transparency are the guiding values that distinguish the IBM approach to AI ethics. The Principles are supported by the Pillars of Trust, our foundational properties for AI ethics. Together, our Principles for Trust and Transparency and Pillars of Trust lay the foundation for how we develop and deploy technology responsibly.

Principles for Trust and Transparency

- The purpose of AI is to augment human intelligence
- Data and insights belong to their creator
- New technology, including AI systems, must be transparent and explainable

Pillars of Trust

- Explainability
- Fairness
- Robustness
- Transparency
- Privacy

Appendix 2: Resources for Developers

IBM Granite Guardian

Granite Guardian 3.0 2B: This model is IBM's smaller-size comprehensive guardian model for detecting social bias and implicit hate, toxicity and explicit hate, profanity, violence, sexual content, unethical behavior, jailbreaking, hallucination/groundedness, context relevance, and answer relevance. It may be used in various ways throughout the LLM lifecycle with any proprietary or open-weight model.

Granite Guardian 3.0 8B: This model is IBM's best-performing comprehensive guardian model for detecting social bias and implicit hate, toxicity and explicit hate, profanity, violence, sexual content, unethical behavior, jailbreaking, hallucination/groundedness, context relevance, and answer relevance. It may be used in many guardrail applications with any proprietary or open-weight model.

[Granite Guardian HAP \(38M\)](#): This model is IBM's lightweight, 4-layer toxicity binary classifier for English. Its latency characteristics make it a suitable guardrail for any large language model. It can also be used for bulk processing of data where high throughput is needed. It has been trained on several benchmark datasets in English, specifically for detecting hateful, abusive, profane and other toxic content in plain text.

[Granite Guardian HAP \(125M\)](#): This model is IBM's 12-layer toxicity binary classifier for English, intended to be used as a guardrail for any large language model. It has been trained on several benchmark datasets in English, specifically for detecting hateful, abusive, profane and other toxic content in plain text.

IBM Granite Models

Check us out on Hugging Face at [IBM Granite!](#)

Granite 2B and 8B Models: A series of lightweight, flexible models for enterprise use, licensed under Apache 2.0 license.

[Granite Code Models](#): A series of code models trained by IBM licensed under Apache 2.0 license. We release both the base pretrained and instruct models.

[Granite Time Series Models](#): A collection of time series models trained by IBM licensed under Apache 2.0 license.

[Granite Geospatial Models](#): A series of geospatial models trained by IBM licensed under Apache 2.0 license.

IBM Opensource Red Teaming Datasets

[AttaQ](#): The AttaQ red teaming dataset, consisting of 1,402 carefully crafted adversarial questions, is designed to evaluate Large Language Models (LLMs) by assessing their tendency to generate harmful or undesirable responses. It may serve as a benchmark to assess the potential harm of responses produced by LLMs. The dataset is categorized into seven distinct classes of questions: deception, discrimination, harmful information, substance abuse, sexual content, personally identifiable information (PII), and violence. Researchers and developers can use this dataset to assess the behavior of LLMs and explore the various factors that influence their responses, ultimately aiming to enhance their harmlessness and ethical usage.

[ProvoQ](#): The ProvoQ dataset is designed to evaluate the sensitivity of large language models (LLMs) to stigma-related topics. It contains 2,705 human-curated provocative questions that systematically target minority-stigma pairs in the United States, creating a diverse and nuanced set of questions that reflect these sensitive topics. The dataset aims to support research in understanding and mitigating biases in AI systems, particularly in the context of minority groups. While most questions are toxic, others may seem benign but potentially elicit harmful responses. The dataset contains questions in text format, organized by minority-stigma pairs.

[SocialStigmaQA](#): Current datasets for unwanted social bias auditing are limited to studying protected demographic features such as race and gender. In this work, we introduce a comprehensive benchmark that is meant to capture the amplification of social bias, via stigmas, in generative language models. We start with a comprehensive list of 93 stigmas documented in social science literature and curate a question-answering (QA) dataset which involves simple social situations. Our benchmark, Social-StigmaQA, contains roughly 10K prompts, with a variety of prompt styles, carefully constructed to systematically test for both social bias and model robustness.

Other Opensource Tools Built on Granite

[IBM Granite Code Installer for Visual Studio](#): IBM Granite Code is an innovative, lightweight AI coding companion built for IBM's state-of-the-art Granite large language models. This companion offers robust, contextually aware AI coding assistance for popular programming languages including Go, C, C++, Java, JavaScript, Python, TypeScript and more. Seamlessly integrated into Visual Studio Code, Granite Code accelerates development productivity and simplifies coding tasks by providing powerful AI support hosted locally on the developer's laptop or workstation using Ollama.

[Granite Speculators](#): A collection of accelerators for the Granite language and code family of models.

[Data Prep Kit](#): Data Prep Kit is a community project to democratize and accelerate unstructured data preparation for LLM app developers. It offers implementations of commonly needed data preparation steps, called *modules* or *transforms*, for both Code and Language modalities, with vision to extend to images, speech and multimodal data. The goal is to offer high-level APIs for developers to quickly get started in working with their data, without needing expertise in the underlying run-times and frameworks.

Appendix 3: Usage Policies and Documentation

IBM Granite models are released under an Apache 2.0 license. For downstream usage of its pre-trained models, IBM makes available the following documentation:

Terms and Conditions: The latest Terms and Conditions for the watsonx platform can be found [here](#).

Product documentation: The IBM Granite models are currently available through IBM's watsonx platform. As part of watsonx, each Granite model is accompanied by a model card that details key facts and provenance of the model.

Technical reports and model cards: Technical reports are publicly released on arxiv.org and may be found on the [IBM Granite website](#). Model cards may be found on [Hugging Face](#).

Responsible Use Guide: The Responsible Use Guide may be found on the [IBM Granite website](#).

Together, this information is designed so that not only IBM complies with legal and ethical requirements, but also to aid the users of the models as they seek to comply with their own obligations.

Appendix 4: Socio-Technical Harms and Risks

Source	Phase	Group	Risk	Indicator
Input	Training and Tuning	Fairness	Bias	Amplified
Input	Training and Tuning	Robustness	False samples	Traditional
Input	Training and Tuning	Value Alignment	Undesirable output for retraining purposes	New
Input	Training and Tuning	Data Laws	Legal restrictions on moving or using data	Traditional
Input	Training and Tuning	Intellectual Property	Copyright and other IP issues with content	Amplified
Input	Training and Tuning	Transparency	Disclose data collected, who has access, how stored, how it will be used	Amplified
Input	Training and Tuning	Privacy	Inclusion or presence of SPI or PII	Traditional
Input	Training and Tuning	Privacy	Provide data subject rights (e.g., opt-out)	Amplified
Input	Inference	Privacy	Disclose PII or SPI as part of prompt to model	New
Input	Inference	Intellectual Property	Disclose copyright or other IP information as part of prompt to model	New
Input	Inference	Robustness	Vulnerabilities to adversarial attacks like evasion (create incorrect model output by modifying data sent to train model)	Amplified
Input	Inference	Robustness	Vulnerabilities to adversarial attacks like prompt injection (force different output), prompt leaking (disclose system prompt), or jailbreaking (avoid guardrails)	New
Output	Inference	Fairness	Bias in generated content	New
Output	Inference	Fairness	Performance disparity across individuals or groups	Traditional
Output	Inference	Intellectual property	Copyright infringement, compliance with open source license agreements	New
Output	Inference	Value alignment	Hallucination (generation of false content)	New
Output	Inference	Value alignment	Toxic, hateful, abusive, and aggressive output	New
Output	Inference	Misuse	Spread disinformation (deliberate creation of misleading information)	Amplified
Output	Inference	Misuse	Generate toxic, hateful, abusive, and aggressive content	New
Output	Inference	Misuse	Nonconsual use of people's likeness (deepfakes)	Amplified
Output	Inference	Misuse	Dangerous use (e.g., creating plans to develop weapons or malware)	New

Source	Phase	Group	Risk	Indicator
Output	Inference	Misuse	Deceptive use of generated content (e.g., intentional nondisclosure of AI generated content)	New
Output	Inference	Harmful code generation	Execution of harmful generated code	New
Output	Inference	Privacy	Expose PI or SPI in generated content	New
Output	Inference	Explainability	Challenges in explaining the generated output	New
Output	Inference	Traceability	Challenges in identifying source and facts for generated output	New
Other	Governance	Transparency	Document data and model details, purpose, potential use and harms	Traditional
Other	Governance	Accountability	Identify responsibility for misaligned output along AI lifecycle and value chain	Amplified
Other	Legal compliance	Intellectual property	Determine creator of downstream models	New
Other	Legal compliance	Intellectual property	Determine creator of open source foundation models	New
Other	Legal compliance	Intellectual property	Determine owner of AI-generated content	New
Other	Legal compliance	Intellectual property	Uncertainty about IP rights related to generated content	New
Other	Legal compliance	Legal uncertainty	Determine downstream obligations	Amplified
Other	Societal impact	Impact on jobs	Human displacement (AI induced job loss)	Amplified
Other	Societal impact	Human dignity	Human exploitation (ghost work in training), poor working conditions, lack of healthcare, unfair compensation	Amplified
Other	Societal impact	Environment	Increased carbon emission (high energy requirements for training and operation)	Amplified
Other	Societal impact	Diversity and inclusion	Homogenizing culture and thoughts	New
Other	Societal impact	Human agency	Misinformation and disinformation generated by foundation models	Amplified
Other	Societal impact	Impact on education	Bypass learning process, plagiarism	New

