Foundation Models and their Use in Software Systems - Trust and Governance

Karthikeyan Natesan Ramamurthy  
Trustworthy Machine Intelligence  
IBM Research AI

NIST Workshop – SSDF for Generative AI and Dual Use FMs  
Secure Use of LLMs and Generative AI Systems
Our leadership in trustworthy AI

Science of Trustworthy AI
foundational works in algorithmic fairness, explainability, robustness, UQ, and transparency.

200+ publications in top AI venues (NeurIPS, AAAI, ICML, ICLR, IJCAI, KDD, CVPR, ICASSP, FAccT, AIES, FSE)

>10,000 citations since 2017
won FICO Explainability Challenge
won VizWiz Challenge
2020 WIRED / HBS Tech Spotlight
2021 WIRED / HBS Tech Spotlight
won Schmidt Futures AI for Good award

Product Contributions
Cloud Pak for Data, Watson Advertising, explainability in MAS-predict, Tririga Insights, SCIS, BTI, Cognos Planning & Analytics, IBM AI Governance

Open Source
leading opensource toolboxes for supporting fair, explainable, and robust AI
AIF360, AIX 360, UQ 360, ART 360
pioneered the concept of FactSheets

Beneficial AI Deployments
Science for Social Good

AI Ecosystem & Policy
PAI, EU Commission High Level Expert Group on AI, NIST, AI Caucus, National AI Strategy, ...

2020 WIRED / HBS Tech Spotlight
2021 WIRED / HBS Tech Spotlight
won Schmdt Futures AI for Good award
IBM’s Approach to Foundation Models & Generative AI for Business

Multi-model

- Build your own models
- Use IBM + open + other models
- Use IBM models

Multi-cloud

Open
Based on the best open technologies available

Trusted
Transparent, responsible, and governed

Targeted
Designed for enterprise and targeted at business domains

Empowering
For value creators, not just users
How can FMs/Generative AI be used in SDLC?

- Methods for software development life cycle (SDLC) as well as the individual software components can be augmented with FMs.

- FM-augmented SDLC techniques includes using FMs for code generation/assistance/review, developing test cases, requirements formulation, design, and documentation.

- Examples of FM-augmented components include using LLMs for generating marketing material, summarizing emails, classifying content as kid-safe and answering user questions based on a knowledge store.

- Trust and governance issues can crop up in both situations and need to be understood and mitigated.

https://insights.sei.cmu.edu/blog/application-of-large-language-models-llms-in-software-engineering-overblown-hype-or-disruptive-change/ [Figure adapted from here]
Examples of FMs/GenAI as components in software systems

- Code generation/documentation for developers – natural language to code, explaining code in natural language.

- Content creation, analysis, paraphrasing, summarization of text/data.

- Search, QA.

- Clustering and classification.

Trust and governance is important in both the individual FM components and for the overall system.

https://www.korbit.ai/post/how-to-build-software-with-llms-part-2 [Figure adapted from here]
What does it take to trust an LLM?

Some AI risks are the **same as in traditional data science**
- poor predictive accuracy
- lack of fairness and equity
- lack of explainability
- model uncertainty
- distribution shifts (drift)
- poisoning attacks
- evasion attacks
- extraction attacks
- inference attacks
- model transparency

But many risks are **entirely new in foundation models**
- hallucinations
- lack of factuality or faithfulness
- lack of source attribution
- privacy leakage
- toxicity, profanities, and hate speech
- bullying and gaslighting
- prompt injection attacks

Occur when LLMs are used in “classical ML” tasks, e.g., prediction and classification, and have well-defined metrics and defenses, i.e. IBM Trust 360 toolkits.

Occur when LLMs are used in generative tasks, and do not yet have well-defined metrics and defenses.
We’ve created a taxonomy of risks to make sure that they are appropriately handled in our technology solutions and governance frameworks.

The range of risks and issues that occur in LLMs is broad, and will be handled in a variety of different ways.

https://www.ibm.com/downloads/cas/E5KE5KRZ
1. Risks associated with input

<table>
<thead>
<tr>
<th>Group</th>
<th>Risk</th>
<th>Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training and tuning phase</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fairness</td>
<td>Bias like historical, representational or measurement bias</td>
<td>Amplified</td>
</tr>
<tr>
<td>Robustness</td>
<td>An adversary or malicious insider injecting false, misleading,</td>
<td>Traditional</td>
</tr>
<tr>
<td></td>
<td>malicious or incorrect samples</td>
<td></td>
</tr>
<tr>
<td>Value alignment</td>
<td>Using undesirable output, such as inaccurate or inappropriate user</td>
<td>New</td>
</tr>
<tr>
<td></td>
<td>content, from downstream applications for retraining purposes</td>
<td></td>
</tr>
<tr>
<td>Data laws</td>
<td>Legal restrictions on moving or using data</td>
<td>Traditional</td>
</tr>
<tr>
<td>Intellectual property</td>
<td>Copyright and other IP issues with the content</td>
<td>Amplified</td>
</tr>
<tr>
<td>Transparency</td>
<td>The ability to disclose what content is collected, how it will be</td>
<td>Amplified</td>
</tr>
<tr>
<td></td>
<td>used and stored, and who has access</td>
<td></td>
</tr>
<tr>
<td>Privacy</td>
<td>Inclusion or presence of personal identifiable information</td>
<td>Traditional</td>
</tr>
<tr>
<td></td>
<td>and sensitive personal information</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Challenges around the ability to provide data subject rights, for</td>
<td>Amplified</td>
</tr>
<tr>
<td></td>
<td>example, opt out, right to access or right to be forgotten</td>
<td></td>
</tr>
<tr>
<td>Inference phase</td>
<td>Privacy</td>
<td>Disclosing personal information or sensitive personal information as a part of prompt sent to the model</td>
</tr>
<tr>
<td>----------------</td>
<td>---------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>Intellectual property</td>
<td>Disclosing copyright information or other IP information as part of the prompt sent to the model</td>
</tr>
<tr>
<td></td>
<td>Robustness</td>
<td>Vulnerabilities to adversarial attacks like evasion, which is an attempt to make a model output incorrect by perturbing the data sent to the trained model</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vulnerabilities to adversarial attacks like prompt injection, which forces a different output to be produced; prompt leaking, which is the disclosure of the system prompt; or jailbreaking, which is avoiding guardrails established in the system prompt</td>
</tr>
</tbody>
</table>
## 2. Risks associated with output

<table>
<thead>
<tr>
<th>Group</th>
<th>Risk</th>
<th>Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fairness</td>
<td>Bias in the generated content</td>
<td>New</td>
</tr>
<tr>
<td></td>
<td>Performance disparity across individuals or groups</td>
<td>Traditional</td>
</tr>
<tr>
<td>Intellectual property</td>
<td>Copyright infringement, including compliance with open-source license agreements</td>
<td>New</td>
</tr>
<tr>
<td>Value alignment</td>
<td>Hallucination—false content generation</td>
<td>New</td>
</tr>
<tr>
<td></td>
<td>Toxic, hateful, abusive and aggressive output</td>
<td>New</td>
</tr>
<tr>
<td>Misuse</td>
<td>Spread disininformation—deliberate creation of misleading information</td>
<td>Amplified</td>
</tr>
<tr>
<td></td>
<td>Generate toxic, hateful, abusive and aggressive content</td>
<td>New</td>
</tr>
<tr>
<td></td>
<td>Nonconsensual use of people’s likeness—deepfakes</td>
<td>Amplified</td>
</tr>
<tr>
<td></td>
<td>Dangerous use—creating plans to develop weapons or malware</td>
<td>New</td>
</tr>
<tr>
<td></td>
<td>Deceptive use of generated content—intentional nondisclosure of AI-generated content</td>
<td>New</td>
</tr>
<tr>
<td>Category</td>
<td>Description</td>
<td>Status</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Harmful code</td>
<td>Execution of harmful generated code</td>
<td>New</td>
</tr>
<tr>
<td>generation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Privacy</td>
<td>Exposing personal information or sensitive personal information in generated content</td>
<td>New</td>
</tr>
<tr>
<td>Explainability</td>
<td>Challenges in explaining why output was generated</td>
<td>Amplified</td>
</tr>
<tr>
<td>Traceability</td>
<td>Challenges in determining original source and facts of the generated output</td>
<td>New</td>
</tr>
</tbody>
</table>
We’ve created a detailed AI risk atlas of 44 harms:
IBM RESEARCH INNOVATION
Risk Assessment
Model Summary View

Snapshot view of the model that provides overall assessment and ongoing monitoring with a breakdown by dimension. Highlights issues and opportunities for investigating the issues by dimension.

1. Model summary overview and details
2. Overall score for the model with breakdown by dimensions
3. Scores by dimension with corresponding threshold
4. Dimension of the model that falls below the predefined threshold
5. Ability to further investigate features of the dimension to understand score
# Metrics for evaluating Large Language Models

## Summarization Metrics
- **Reference based Metrics**
  - From Hugging Face Evaluate Package
    - ROUGE - Rouge 1, Rouge 2, Rouge L, Rouge LSUM
  - SARI
  - Text Quality
  - Normalized F1, Precision, Recall
  - METEOR
  - BLEU
- From OpenSource
  - Sentence Similarity
    - Jaccard Similarity
    - Cosine Similarity
  - Levenshtein distance based Diversity metrics
- **Reference-free Metrics**
  - From IBM Research
    - HAP Detection
    - PII Detection
  - From Open Source
    - Readability, complexity
    - Blanc

## Entity Extraction Metrics
(Deterministic data extraction, Contextual text extraction – example contract clause)
- From Hugging Face Evaluate Package
  - Seq eval
- From IBM Research Suggested Metrics
  - Micro & Macro F1, Precision, Recall

## Content Generation Metrics
- From Hugging Face Evaluate Package
  - ROUGE - ROUGE 1, ROUGE 2, ROUGE L, ROUGE LSUM
  - BLEU
  - METEOR
  - exact_match
- From Open Source
  - ROUGE
  - From IBM Research
    - HAP Detection
    - PII Detection

## Q&A Metrics
(RAG – Retrieval Augmented Generation = Search & Summarize)
- From Hugging Face Evaluate Package
  - ROUGE - ROUGE 1, ROUGE 2, ROUGE L, ROUGE LSUM
  - BLEU
  - METEOR
  - exact_match
- From Open Source
  - ROUGE

## Explainability Monitoring
- Attribution - IBM Research’s alternative to cosine similarity

## Classification Metrics
- Metrics that OpenScale already monitors for Text Classification
  - Accuracy
  - Precision
  - Recall
  - ROC AUC
  - F1 Score
- From Hugging Face Evaluate Package
  - Brier Score
  - Matthews Correlation Coefficient
  - Label Skew

## Fairness/Bias Monitoring
- Protected Attributes Extraction on the prompt output and evaluate Fairness on Classification output
- Fairness evaluation when fairness attributes are logged as meta attributes via., Payload/Feedback Logging

## Drift Monitoring - OpenScale specific algorithms
- Structure Drift
- Content Drift
- Confidence Drift
- Distribution Drift
- Root Cause Analysis
Traditional AI to Generative AI
Trustworthy & safe foundation model lifecycle for enterprise FM governance

General-purpose
dialogue, academic, internet, code, books, ...

Domain specific
Legal, Financial, Regulatory, Medical, ...

Proprietary
Enterprise specific, e.g., transactions, network data
IBM Data Pile

Data inspection, curation, and cleaning
toxicity detection, removal, hate & profanity filtering, augmentation

Architecture selection
encoder only, encoder-decoder, decoder only ...
compute-optimal scaling, quantization friendly models, novel sparse architectures, spiking architectures, architectures with memory

Training
private training, efficient progressive training, quantization down to 4-bit, sparse training, bias-aware training, training with trust constraints

Model Adaptation
fine-tuning
reprogramming
prompt engineering, prompt tuning, prompt design

Model Teaching, & Improvement
fine-tuning for trust
reprogramming for trust
prompt-tuning for trust
learning from human feedback
learning from AI feedback
Principled AI: Alignment studio
post-processing

Governance
usage guidance
risk assessment
fact collection (FactSheets)
model audit
policy packs
model monitoring & safeguards

points of “Trust instrumentation” and governance
The Foundation model journey: from training to usage

Foundation model provider

- **Pre-trained from scratch**
  - Train a custom foundation model from scratch. (100’s GB to 10s TB)

- **Custom training**
  - Modifies all the underlying weights of an underlying basic FM (10’s GB)

- **Full fine-tuning**
  - Reinforcement learning
  - Modify some of the underlying weights to adapt model behavior and safety properties (100’s-100000’s of examples)

Foundation model consumer

- **Parameter efficient fine-tuning**
  - Create a small adaptor while keeping the underlying model frozen. (100’s-1000’s of examples)

- **Prompt-engineering**
  - Providing few shot text examples in the prompt. (10’s of examples)

- **Post-processing**
  - Modify the output (1)

Data/complexity / cost

Model and data

Modify the model parameters

Add parameters

Modify the prompt

Modify the output
Data Governance underlying IBM models

Data collection and extraction

External Datasets

IBM Internal Datasets

Data cleaning

- Document id generation
- Exact dedup
- Fuzzy dedup (threshold)
- Language detection
- Sentence splitting

Data annotation

- Hate, Abuse, Profanity filters
- License filters
- PII filters
- ...

Data filtering

Tokenization

FM training

Dataset Acquisition

Dataset Preprocessing (model neutral)

Data Preprocessing (model specific)
Principled AI: The Mitigators

Novel safeguards, guardrails, and other correction mechanisms

- **fine-tuning, prompt-tuning, reprogramming, and post-processing for bias correction**
  - “Equi-Tuning: Group Equivariant Fine-Tuning of Pretrained Models,” AAAI 2023
  - ”Fair Infinitesimal Jackknife: Mitigating the Influence of Biased Training Data Points Without Refitting,” NeurIPS 2022
  - “Fairness Reprogramming,” NeurIPS 2022
  - “Post-processing for Individual Fairness,” NeurIPS 2021

- **quantifying uncertainty in model outputs**
  - “Learning Prediction Intervals for Model Performance,” AAAI 2021

- **explaining model outputs**
  - “Let the CAT out of the bag: Contrastive Attributed explanations for Text,” ACL 2022

- **detecting generated text**
  - “GLTR: Statistical detection and visualization of generated text,” ACL 2019

- **measuring faithfulness**

- **detecting undesirable behaviors**
  - “Detecting Egregious Conversations between Customers and Virtual Agents,” NAACL 2018

- **privacy preservation**
  - “Reprogrammable-FL: Improving Utility-Privacy Tradeoff in Federated Learning via Model Reprogramming,” IEEE SaTML 2023
Generic harms vs. specific harms

Common across sectors and use cases

Unique or particular to a company

- laws
- industry standards
- social norms of end-users
- corporate policies
- market demands
- technology architecture constraints

Alignment approaches are too generic and cannot be controlled

Open AI
InstructGPT

Anthropic
Constitutional AI

Meta
Llama-2-Chat

Principled AI alignment studio
For the entire software lifecycle with FM components

- Trust and governance is critical for each FM component in the software system.

- It is important also to ensure that the entire software system is governed end-to-end and subjected to risk assessment and mitigation.

- Trust and governance for the entire software system with FM components has not been subject to rigorous study yet. However, even for standalone LLMs, adversarial fine-tuning has been shown to break alignment with a handful of instances.

- The first step toward trust and governance for the entire software system is to understand the risks and develop benchmarks for quantifying the risks.

- The next step is to develop ways for mitigating the risks.

- Some of the existing risk and mitigation measures developed for FMs could be used for the entire software system also.

Take-home messages

- Use of FMs in general and LLMs in particular is very promising in existing software systems.

- They can be used as components (FM-augmented systems) or can be used to guide the SDLC (FM-augmented SDLC) or in both (FM augmented systems built with FM augmented SDLC).

- Trust challenges are similar in all these cases.

- Ensuring that the individual FM components are trustworthy is necessary but not sufficient since downstream and upstream components can still make the software system un-trustworthy.

- End-to-end assessment of trust is critical – need to understand, quantify, and mitigate risks.