DATA LANDSCAPE of GEN AI for Project Managers

Risks and Limitations of Data in GenAI

WHO TO TALK TO AND WHAT TO ASK

When using data and GenAI on an enterprise level within projects, project managers need to have a high-level understanding of these concepts and discuss them with experts in the organization. Join forces with data teams, security teams, and other experts to manage risks. It is the project manager's role and responsibility to ensure these conversations take place and to facilitate them.

This document outlines thirteen categories of risk, what project managers should do, and suggested roles and questions to ask experts to manage these risks, protect sensitive data, maintain legal compliance, and ensure that projects uphold the highest ethical standards. Project managers should to have these conversations and checkpoints with your experts regularly—this is an iterative process.

The suggested task areas and role accountability in this document will vary from organization to organization and will be dependent on the data governance policies, processes, and programs in place. If these are not in place, the project manager could work with the governance and other teams to establish them.

TASK AREA AND ROLE ACCOUNTABILITY

	Product Owner	AI Team/Data Scientist	Data Architect/Solution Architect	Privacy & Compliance Lead	Legal Department	Ethics Committee	IT Security	Data Owner/SME	Data Governance Team	IT Support
Data Privacy				✓	✓	✓	✓			
Biased Data & Uncertain Outputs					✓	✓		✓		
Data Ownership & Control					✓			✓	✓	
Data Minimization		✓	✓					✓		
Transparency & Consent					✓			✓	✓	
Data Retention & Deletion	✓	✓							✓	✓
Monitoring & Compliance				✓	✓	✓	✓			
Lack of Explainability		✓						✓		
Dependence & Over-reliance		✓						✓		
Generalization and Overfitting		✓						✓		
AI Hallucinations		✓						✓		
Model Collapse and Creativity		✓						✓		
Uncertain Future and Evolving Nature		✓						✓		

DATA PRIVACY

Training GenAI can involve large datasets whose collection, storage, and management can imply privacy risks. With sensitive project data at stake, unauthorized access or misuse can have severe legal and ethical implications. A breach exposing confidential information can erode stakeholder trust and lead to legal issues.

It's your job as the project manager to ensure that data used in the GenAI training process complies with data protection regulations, respects user privacy, and is appropriately secured.

What roles should you reach out to?	What questions should you ask?
Privacy and Compliance Lead	**After describing the project-related data and how it will be used** • May I have a copy of any privacy policies that pertain to this project? • Which aspects of the privacy policy should I focus on and why? • Could you describe the intent behind those aspects of the policy? • Which regulations apply to the project data I've described? • How do we typically mitigate privacy risk with this type of data?
Legal Department	**After describing the project-related data and how it will be used** • What privacy laws and regulations might be of concern, given the scope of our project and the data sets it touches? • How do we typically mitigate privacy risk with this type of data? • Are there any permissions, copyright, or license agreements we need to seek to use the data I've described? • Any security concerns related to this data being used to generate potentially derivative works? • How should we think about ownership and other legal issues related to the data and information that will be generated by this new system?
Ethics Committee	**After describing the project-related data and how it will be used** • Do you have any ethical concerns about the use of the data I've described? • Any ethical concerns about the broader effort we are undertaking? • Any advice as this system evolves over time?
IT Security Team	**After describing the project-related data and how it will be used** • Where is this data currently stored? Would this be a separate instance? • Will making it accessible to our project cause any specific security concerns? • Should the data be anonymized when stored? • How should we think about restricting permissions? • How should we think about security related to the data and information that will be generated by this new system?

BIASED DATA AND UNINTENDED OUTPUTS

GenAI systems rely heavily on data and biases present in the data can lead to biased GenAI outputs. Poor quality or unrepresentative data can result in inaccurate or unfair conclusions.

You must be vigilant to prevent the output of biased, discriminatory, or offensive data that can harm the project's reputation and legal standing. Implementing ethical guidelines and reviewing generated content is essential.

What roles should you reach out to?	What questions should you ask?

Data Owner/SME	**Test the system early and often with the Data Owner/SME**
	• Demonstrate typical prompts and responses.
	• What biases have you uncovered in the data this system will use?
	  • Does any of the data reflect historical or social inequities, or include sensitive variables such as gender, race, or sexual orientation?
	  • Any known behavioral biases reflected in the data?
	  • Are any groups over or underrepresented in this data set?
	• What is the original source for this data? Is it reliable?
	• How do you address bias and fairness today with respect to this data?
	• What safeguards are already in place?
	• Do we need any new policies related to data usage for existing data or generated information?

Legal Department	**Review permissions, data being used, and types of information likely to be generated** • Is everyone who will have permission to use this system clear on their legal responsibilities relative to both the data being used and the information generated? • What legal exposure might be created by biased or incorrect information generated by this system? (ensure these are feasible use cases) • Do we need any new policies related to data usage for existing data or generated information? • Are the proper references available to avoid copyright violations?
Ethics Committee	**Test the system early and often with the Ethics Committee** • Demonstrate typical prompts and responses. • What biases have you uncovered in our systems recently and would those apply here?

DATA OWNERSHIP AND CONTROL

Data, content, and images generated by GenAI may carry associated risks concerning their ownership and rights.

It's your responsibility to clarify data ownership and control when using GenAI to create content. This requires transparent data usage policies.

TRANSPARENCY AND CONSENT

Gaining user consent before using their data for GenAI training is crucial.

Make transparency and user consent a priority. Users should be informed about the use of GenAI and its potential impact on their data.

The following roles and questions relate to both data ownership and control, and transparency and consent.

What roles should you reach out to?	What questions should you ask?

Data Governance Team	**After describing the project-related data and how it will be used**
	• What ownership responsibility for data, being consumed or generated for this project, has been put in place?
	  • Notify the appropriate data owners regarding all downstream changes in data usage
	  • Note: Data ownership should follow an organizational framework. The owners of a particular area of data should be aware of when and how the data they own will be used.

Data Owner/SME	**After describing the project-related data and how it will be used** • How is the data associated with your data domain, and used in this project, collected? Where did it originate? • How is it stored? • How often is it updated? • What refresh rate should be used for data that feeds the GenAI • application?
Legal Team	**After describing the project-related data and how it will be used** • Are there any legal issues with permissions or licensing of any of the data sets being used or generated? • Do any user agreements need to be updated to reflect the new ways this data will be used? • Will there be any ownership issues or challenges with usage related to data or information generated?

DATA MINIMIZATION

The vast amounts of data used in GenAI are only sometimes required to accomplish the intended tasks.

Ensure data collected and used by the GenAI is limited to only what's required for the project.

What roles should you reach out to?	What questions should you ask?
AI Team/Data Scientist	• How often will training take place? • How can we minimize the amount of data used through better data quality or improvements in algorithms/models? • Will duplicate inputs be reviewed? • Once the tool is trained, is the source data still needed? • For how long? Note: Data ownership should follow a framework within the organization the owners of a particular area of data should be aware of when and how the data they own will be used.
Data Architecture/ Solutions team	• What is our assigned storage capacity? • How long should data be held?
Data Governance	• What is our data retention policy globally? • What is our data retention policy for the data elements related to this project? • What is our data retention policy for the data this project will generate?

DATA RETENTION AND DELETION

Projects that incorporate GenAI should define clear data retention and deletion policies.

Ensure that data is retained only as necessary and is deleted adequately upon request.

MONITORING AND COMPLIANCE

Ongoing monitoring and compliance with data protection regulations are essential.

Collaborate with legal and compliance experts to verify that use of GenAI adheres to all relevant laws and standards.

The following roles and questions relate to both data retention and delection, and monitoring and compliance.

What roles should you reach out to?	What questions should you ask?
Data Governance	• Who monitors and polices compliance with the data retention policy?
IT Support	• What is the data removal process? • Is there archiving capability?
Project Owner	• Is there any data that should be held for further analysis after retention timeframe is hit? • If yes, what data should be kept?
Data Science Team	• How will user input, chat questions, be used to continue to train the model? • How often should user input be reviewed and applied? • When should this data be removed?

LACK OF EXPLAINABILITY

Many GenAI models are often seen as black boxes, making it difficult to understand how they arrived at a particular decision or prediction. This lack of transparency can be problematic in situations where explainability is crucial.

Ensure that data is retained only as necessary and is deleted adequately upon request.

GENERALIZATION AND OVERFITTING

GenAI models might “overfit” to the training data and fail to generalize well to new, unseen data. This means they perform well on data they’ve been trained on but poorly on new data.

Collaborate with data scientists to utilize diverse and extensive datasets for training to enhance generalization and regularly validate the GenAI model against new and unseen data.

DEPENDENCE AND OVER-RELIANCE

More reliance on GenAI can lead to a decline in human decision-making skills. There’s also the risk of becoming too dependent on specific tools or vendors. The recent uncertainties with GenAI vendors like OpenAI underscore the risks of over-reliance on external GenAI solutions.

Balance dependence on these tools while being prepared for potential disruptions.

AI HALLUCINATIONS

AI, particularly in generative models, can ‘hallucinate’ or generate false or nonsensical outputs, which can be misleading.

Implement rigorous testing and validation processes for generative models and monitor model outputs for accuracy and consistency.

MODEL COLLAPSE AND CREATIVITY

An emerging concern is model collapse, where AI trained on AI-generated data starts losing creativity and originality, echoing only what it has seen during training. This phenomenon can limit GenAI's innovative potential.

Regularly update and diversify GenAI training data to prevent model collapse. Monitor models for signs of model collapse. Foster a culture of continuous improvement and innovation in GenAI development.

UNCERTAIN FUTURE AND EVOLVING NATURE

The GenAI field is rapidly evolving, making it difficult to fully predict its trajectory and implications. This uncertainty can be a risk for long-term planning and investment.

Stay updated with the latest AI trends and developments and adapt project goals and strategies per evolving AI capabilities. Lead the team in agile adaptation to the changing AI landscape.

The following roles and questions relate to lack of explainability, generalization and overfitting, dependence and over-reliance, AI hallucinations, model collapse and creativity, and uncertain future and evolving nature.

What roles should you reach out to?	**What questions should you ask?**
Data Science Team	• Does the model meet our expectations? • Are there new models, approaches, or combinations of models that we should be considering? • Are there new packaged solutions that now address our major project challenges? • What questions related to our approach, should we be asking at this juncture?
Data Owner/SME	• Is our approach the best way you know of to get the answers we are seeking? • Is it performing as you would expect? Why? • What other ways are people doing this today?