How to assess privacy and reliability before using generative AI

Why assessment matters before you trust generative AI

Generative AI tools can produce text, images and code in seconds. But the quality of what they produce depends heavily on how they were built, what data they were trained on, and how they are deployed. Before you use any generative AI application in a professional setting, you need a clear method to judge whether it respects privacy and whether its outputs can be trusted. This is especially true when the tool will handle sensitive information or support decisions that affect people's work or lives.

Step 1: Check what data the model was trained on

The first question to ask is about the training data. A model that was trained on public internet data may contain personal information, copyrighted material, or biased content. If you are using a tool that was fine-tuned on a specific dataset, ask the provider what that dataset included. Look for documentation that explains the sources, how consent was obtained, and whether any personal data was removed before training. Without this, you cannot know what the model might reproduce or how it might skew its outputs.



Watch for vague statements like "trained on diverse data" without specifics. Responsible providers publish a data sheet or a model card that lists the training data's origin, size, and any filtering steps. If that information is missing, treat the tool with caution until you can verify its background.

Step 2: Understand how your inputs will be handled

When you type a prompt or upload a file, that data leaves your device and travels to the model provider's servers. You need to know what happens to it next. Does the provider log your prompts? Do they use them to retrain or improve the model? Do they share them with third parties? These details should be in the provider's privacy policy or terms of service.

Look for clear commitments: data is encrypted in transit and at rest, it is not used for training unless you opt in, and it is deleted after a defined period. If the tool allows you to run the model locally on your own hardware, that gives you more control. For cloud-based tools, check whether the provider has a data processing agreement that complies with regulations like Singapore's Personal Data Protection Act (PDPA) or the GDPR if your work crosses borders.

 How to assess privacy and reliability before using generative AI

Step 3: Evaluate the model for bias and fairness

Generative AI models can reflect and amplify biases present in their training data. A model that is used for hiring, customer service, or content moderation needs to be tested for fairness across different demographics. Look for bias evaluation reports from the model developer. These should show how the model performs on tasks that involve race, gender, age, language, and other sensitive attributes.

If no such report exists, run your own small tests. Give the model prompts that vary only by a demographic detail and see if the outputs differ in tone, accuracy, or completeness. Be aware that a model may also produce harmful stereotypes or offensive language. A reliable provider will have content filters and fine-tuning to reduce these risks, but no filter is perfect.

Step 4: Set up human oversight and access control

Even the most reliable model can make mistakes. Every generative AI application needs a human in the loop who reviews outputs before they are used. Decide who gets to see what the model produces. Not everyone in an organisation should have the same level of access. For example, a junior employee might use the tool to draft content, but a manager should approve the final version. Access control also means limiting who can feed sensitive data into the model. If the tool is used for customer support, ensure that personal identifiers like names and account numbers are stripped from prompts before they are sent.

Document these controls in a simple policy. That policy should say what kinds of queries are allowed, who can approve outputs, and how to report problems. This is not just about safety: it also builds trust with customers and regulators who want to see that you are managing the technology responsibly.

Step 5: Monitor outputs and log issues

Once the tool is in use, you need ongoing monitoring. Output monitoring means checking a sample of the model's responses for accuracy, appropriateness, and any signs of data leakage. If a model accidentally reveals a piece of personal information from its training data, that is a serious privacy breach that needs immediate action.

 How to assess privacy and reliability before using generative AI

Set up a log that records each query, the model's response, and the human reviewer's decision. This log helps you spot patterns: maybe the model struggles with certain topics, or it starts producing odd responses after a software update. Regular reviews of these logs allow you to adjust your prompts, retrain the model, or switch to a different provider if needed.

Step 6: Perform a risk assessment before deployment

A formal risk assessment brings all these steps together. Write down what could go wrong: the model could leak sensitive data, produce biased results, or generate incorrect information that leads to a bad decision. For each risk, rate its likelihood and its potential impact. Then list the controls you have put in place to reduce that risk. This document becomes the basis for deciding whether to deploy the tool and under what conditions.

The risk assessment should be reviewed by someone who is not the same person who set up the tool. An independent pair of eyes often catches gaps that the implementer missed. If the tool will handle particularly sensitive information, consider involving a legal or compliance officer.

Where to learn more about responsible AI practices

If you are looking for structured guidance on building and deploying generative AI responsibly, resources like the workforcelearning.unitedceres.edu.sg course cover exactly this territory. The program is designed for professionals who want to move from theory to practice in two days, with hands-on work on a simple generative AI application. That kind of focused learning helps you internalise the assessment steps described here, so you can apply them in your own organisation.



How to assess privacy and reliability before using generative AI

Understanding how to assess privacy and reliability is not a one-time task. As models improve and new providers enter the market, your criteria should evolve. The same principles that protect your data today will also guide you toward tools that are transparent, fair, and accountable. Keep your checklist handy, revisit it regularly, and never skip the human review step. The technology is powerful, but it is only as trustworthy as the practices you put around it.

For a deeper dive into these topics, the **workforcelearning.unitedceres.edu.sg** course offers a practical route to building your own generative AI application while embedding these checks from the start. Whether you are a manager evaluating a vendor or a developer integrating a model, the skills you gain there will help you ask the right questions at every stage.