MODULE 6 · STUDY NOTES

Practicing Generative AI Responsibly

Responsible-AI dimensions, generative AI risks, safeguards across the workflow, and the AWS tooling that implements them

The creative, open-ended power that makes generative AI attractive is exactly what makes it risky: the same model that drafts a story can invent citations, echo a bias, or reproduce protected content. This module builds the framework for managing that risk. It starts by defining responsible AI and AWS's eight core dimensions of it, examines the trade-offs among transparency, interpretability, and explainability, then names the four challenges generative AI amplifies — hallucinations, toxicity, bias, and illegality. From there it maps safeguards onto every stage of a generative AI application, from prompt to output to production monitoring, and finishes with the AWS tools that implement those safeguards: SageMaker Model Cards, Clarify, Ground Truth, and A2I on one side, and Amazon Bedrock's evaluations and Guardrails on the other.

LEARNING OBJECTIVES

- Define the core dimensions of responsible AI
- Discuss the challenges of generative AI
- Discuss the trade-offs of transparency, interpretability, and explainability
- Identify AWS services and tools that support building responsible AI solutions

6.1 Defining responsible AI

Responsible AI is an approach to building and using AI systems that is **people-centric, ethical, safe, and unbiased**. *People-centric* extends human-centered design — focusing on user needs and behaviors — with practices aimed at generative AI's specific risks, and it reserves time for **human oversight across all stages** of development and in production. *Ethical* means weighing the moral implications of AI and its impact on society: fairness, human rights, social good. *Safe* means systems that are reliable and robust and pose no risk to users or the environment. *Unbiased* means preventing and mitigating unfair discrimination through careful data selection, algorithm design, and continuous monitoring of outputs. These aren't bolt-ons: responsible AI integrates across the whole lifecycle — data selection and preprocessing, model design and training, testing and validation, and ongoing monitoring and adjustment of deployed systems.

TABLE 6.1 AWS core dimensions of responsible AI

DIMENSION	FOCUS	RESPONSIBLE PRACTICES
Fairness	Minimize unequitable impacts on subpopulations (gender, ethnicity, age, language)	Diverse, representative training data · detect + mitigate bias · fairness metrics · regular audits · inclusive design
Safety	Design and test systems to avoid unintended harm to humans or the environment	Guardrails · risk assessments · plan for misuse or failure · keep goals aligned with human intentions
Privacy & security	Protect data from theft and exposure; individuals control if and when their data is used	Granular access controls · data encryption · anonymize sensitive data · regular audits
Veracity & robustness	Correct outputs even with unexpected or adversarial inputs	Adversarial training · ensemble methods · regular evaluation · keep models up to date
Transparency	Understand how the model or system works; openness about capabilities, limits, risks	Documentation · data transparency · UI disclosure of AI use · model cards · auditability
Explainability	Clear, understandable explanation of how a particular output was reached	Model-agnostic explanation techniques (SHAP, LIME, and others in §6.3)
Controllability	Make sure AI systems align with human values and intent	Clear mechanisms for human intervention and adjustment · monitoring + transparent real-time reporting
Governance	Framework so AI is developed and deployed to legal, ethical, and societal standards	Clear policies and guidelines · oversight mechanisms (ethics committees, audits) · compliance with laws and best practices

THREE GROUPS, ONE FRAMEWORK

- **Fairness, safety, privacy & security, veracity & robustness** protect the users of the system and its data.
- **Transparency and explainability** make the system's work open and understandable — the basis of trust and a mechanism for finding bias.
- **Controllability and governance** keep humans actively managing AI behavior and make responsible practice part of how the organization operates.
- The dimensions reinforce each other: better explainability can improve fairness by revealing bias; strong governance keeps privacy and security consistently applied.

Fairness in practice. Fairness is typically measured by analyzing bias in a model's predictions across different groups, hunting for disparities in performance or decision-making. Training data is often the **primary source of bias**, so diverse and representative datasets — with careful collection, preprocessing, and augmentation — are the first defense. Fairness metrics quantify equity across subgroups: examples are **demographic parity (DP)**, **predictive parity (PP)**, and **conditional use accuracy equality (CUAE)**, each ranging **0–1** (expressed as a decimal or percentage). For foundation models a common approach is **SHAP** (Shapley additive explanations), which computes each feature's contribution to a prediction — explaining decisions and surfacing bias at the same time. Round it out with ongoing audits of outputs across demographic groups and inclusive design that brings diverse stakeholders into development.

TABLE 6.2 Types of bias that responsible AI addresses

BIAS TYPE	HOW IT ARISES
Data bias	Training data underrepresents certain groups, skewing model outputs
Algorithm bias	The algorithm itself exhibits prejudiced patterns — via proxy discrimination or optimization metrics that favor performance over fairness — even with representative data
Interaction bias	Skewed or unrepresentative human interaction and feedback during development or real-world use leaks into outputs and future iterations
Bias amplification	Existing societal biases are learned, perpetuated, and magnified by the system

Transparency, interpretability, explainability. *Transparency* is the umbrella: the degree to which you can understand how an AI model or system works. It's built from clear communication and documentation, openness about data sources and their biases, user interfaces that disclose when AI is in use, algorithm transparency balanced against intellectual-property protection, auditability by third parties, and **model cards** — standardized documentation of a trained model's performance characteristics and intended use cases. Within that umbrella sit two distinct ideas:

INTERPRETABILITY	EXPLAINABILITY
understand the mechanics	*explain the outcome*
▪ Deep understanding of the model's internal workings — algorithms, data, decision-making process ▪ Highly interpretable models (e.g. decision trees) are highly transparent ▪ Comes with simpler, less complex models	▪ High-level, human-terms account of how a particular output was reached ▪ Model-agnostic techniques explain opaque-box models that aren't interpretable ▪ The practical route for complex models, including FMs

Explainability matters for four reasons: **trust and transparency** (people adopt what they can understand), **regulatory compliance** (many industries require it for high-stakes decisions), **debugging and improvement** (understanding decisions exposes issues, biases, and limitations), and **ethics** (assessing fairness in decision-making). Which to pursue depends on the use case and the level of transparency it requires.

TRANSPARENCY TRADE-OFFS

- **Accuracy vs. transparency** — complex, highly accurate models are less interpretable than simple ones; interpretable models (decision trees) sit at low complexity, explainability techniques serve the high-complexity end.
- **Privacy vs. transparency** — protecting data privacy can limit the detail in model explanations.
- **Safety vs. transparency** — safety measures can reduce visibility into the model's raw reasoning.
- **Security vs. transparency** — highly secure training environments can limit external auditing.
- Transparency also carries risks: higher costs, compromised competitive edge (proprietary details exposed), privacy/security exposure, unrealistic user expectations of perfect explanations, and model vulnerabilities that malicious actors can exploit.

COMMON PITFALL

Keep the vocabulary straight: **interpretability** is understanding the model's internals; **explainability** is a human-terms account of one output, usually produced by model-agnostic techniques applied to an opaque-box model. A middle option exists too — open-source models or systems with detailed documentation of data sources, preprocessing, and architecture offer *medium* transparency at medium complexity.

Why business cares. Responsible AI pays beyond ethics: **increased trust and reputation** (stronger relationships with customers, partners, stakeholders), **improved regulatory compliance** (staying ahead of AI-specific regulation), **improved risk mitigation** (catching bias, privacy violations, and security breaches before they cause financial and reputational damage), **competitive advantage** (differentiated products that attract customers and talent), **enhanced decision-making** (fair, transparent, accountable AI yields more informed decisions), and **improved products and processes** (diverse perspectives produce more inclusive, effective solutions).

6.2 Generative AI considerations: four challenges

Foundation models are *creative* — they predict plausible output rather than retrieve verified facts — and that open-endedness produces four characteristic risks: **hallucinations, toxicity, bias, and illegality**. Techniques exist to mitigate each, but **an inherent risk always remains**. The governing rule: if a use case has no tolerance for these risks, don't use a generative approach for it.

TABLE 6.3 Generative AI challenges – examples, considerations, safeguards

CHALLENGE	EXAMPLES	CONSIDERATIONS	SAFEGUARDS
Hallucinations	Inaccurate information; fictional citations; weird or creepy generated images	Tolerance differs by use case; fact-checking and human oversight are critical	Fact-checking mechanisms · deterministic inference parameters · data augmentation for relevancy · HITL validation
Toxicity	Discriminatory language; sexually explicit material; advice with potentially negative impacts	Defining rules and boundaries is critical; definitions vary by use case; balance safety against censorship	Policies and guidelines · system guidance disallowing unsafe topics · exclusions and filters · human oversight
Bias	Job descriptions assuming a male candidate; image galleries lacking representation; recommendations catering to higher incomes	Mitigation matters more when training data is unknown; feedback loops amplify bias; transparency + explainability are key	Bias-mitigation techniques · regular audits · continuous monitoring and feedback loops · human review and intervention
Illegality	Synthetic voice imitating a celebrity without permission; copyrighted passages in output; exposed training data breaking privacy law	Disclose system capabilities and limitations; curate data carefully; robust content filtering required	Automated plagiarism and IP checks · model auditing/testing for leaks · HITL with subject matter experts and legal professionals

Hallucinations follow from the model's creative nature: assertions that *sound* plausible but are verifiably incorrect — a common LLM phenomenon is inventing **nonexistent scientific citations**. Tolerance is a design decision: creative storytelling absorbs a hallucination that a **legal briefing** cannot. **Toxicity** is output that is *hateful, threatening, insulting,*

or demeaning to an individual or group. A model is considered safe when output excludes content that could negatively impact individuals or society — including specific individual medical, legal, political, or financial advice, and advice on building weapons. The hard part is the boundary: is an offensive quotation acceptable when clearly labeled as a quotation? An offensive opinion labeled as opinion? The answer depends on audience and context — a children's book has no room for vulgarity — and identifying subtly or indirectly worded offense is a real technical challenge.

Bias enters when the model was trained on biased data or learned biases during training; generating new content from that data perpetuates and can amplify it. With pre-trained FMs you can't see the training data, so bias mitigation during evaluation and output testing becomes critical. **Illegality** covers output containing intellectual property or private data present in the training data — direct or closely imitated use of protected text, code, or imagery. The boundaries are still being defined (when does *in the style of* an artist become infringement? a celebrity's voice was recently imitated without permission), copyright holders are filing suits, and legislation is catching up. Exposure of **personally identifiable information (PII)** is both an ethical and a legal risk. Addressing IP concerns will take a combination of technology, policy, and legal mechanisms over time.

> **COMMON PITFALL**
>
> Exam questions test the *posture*, not just the list: hallucinations, toxicity, bias, and illegality can be **managed, not eliminated**. The common safeguards across all four are human oversight, robust monitoring and auditing, and output filtering — and the fallback is choosing a non-generative approach when tolerance is zero.

6.3 Implementing safeguards across the application

Safeguards aren't a single component — they map onto every stage of the generative AI flow (*user* → *prompt* → *FM* → *output* → *user*). At the **prompt**: prompt engineering, data augmentation, and curation. At the **FM**: model transparency, explainability, and evaluation. At the **output**: content filtering and human-in-the-loop reviews. Around the loop: **monitoring, auditing, and feedback loops**. **Data protection extends across the entire flow**, and responsible-AI guidelines, policies, and processes sit on top of everything.

> **SAFETY FIRST: DESIGN FOR RESPONSIBLE AI**
>
> - Incorporate AI governance into the organization's AI strategy **from the outset** — make responsible AI core to project requirements, not an afterthought.
> - Implement or develop guidelines at the **start** of each project; develop guidelines, policies, and processes **iteratively** — you won't write a policy once and shelve it.
> - Expect to adjust as technology, capabilities, and attitudes evolve; involve everyone in the process.
> - Educate users about how generative AI actually works: **it's not a search — it predicts the right output based on its training**.
> - The AWS whitepaper *AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI* includes a governance-for-responsible-AI section.

Effective prompt engineering *is* responsible AI. Prompt elements guard against risk directly: **specify tone and audience** (telling the model the output is for youths controls toxicity), use **exclusions or negative prompts** to keep certain words or output types out, give **few-shot examples of appropriate output** to keep the model aligned with intent and away from hallucinations, and **provide input data** to raise the relevancy of the output.

TAKING CARE OF DATA

- Provide additional input data via **RAG, fine-tuning, or continued pre-training** — e.g. RAG over internal medical references, or over a verified citation database to mitigate hallucinations.
- **Carefully curate the data sources you control**: input prompts, RAG data, and model-customization data. If you build your own FM, curating training data helps mitigate toxicity and unfairness at the source.
- With an existing pre-trained FM you can't control training data — so shift focus to the **outputs**: avoid toxicity, mitigate bias, prevent IP and private-data leaks.
- **Encrypt data in transit and at rest end to end** — preparation, training, and inference — for privacy, compliance, and breach protection. Generative apps often touch highly sensitive personal, compliance, operational, and financial data.
- Apply **comprehensive, granular access control** to the application and each component; if you fine-tune or build models, secure the model itself and its training and deployment configurations.

Explainability strategies. During training, *features* — characteristics of the data — emerge as important to predictions and are weighted differently. In interpretable models those weights are easy to see and adjust; in complex models (including FMs) there is no direct view of how each feature influences a prediction. **Model-agnostic** techniques fill the gap, working with any ML model to build trust, assist debugging, and mitigate bias — for example, if analysis shows a text-generation model leaning heavily on gender-specific terms, that flags potential gender bias, prompting a configuration tweak, better data, or a different FM.

TABLE 6.4 Explainability strategies for responsible AI

TECHNIQUE	WHAT IT DOES
Feature importance analysis	Quantifies the contribution of each input feature to the model's predictions
Partial dependence plots (PDPs)	Graphical visualization of how changing one specific feature affects predictions while all other features stay constant
Shapley additive explanations (SHAP)	Explains the output of any ML model by calculating each feature's contribution to the prediction
Local interpretable model-agnostic explanations (LIME)	Creates a simplified version of the model around one specific prediction, then explains the prediction by interpreting the simpler model
Ask the LLM to explain itself	Include a justification request in the prompt — e.g. classify an email into one of four categories *and justify the classification* — and the LLM narrates its reasoning in natural language

COMMON PITFALL

Explainability tools are powerful but **not perfect** — they can produce inconsistent or misleading explanations, especially for very complex models. Use them inside a broader responsible-AI strategy with diverse perspectives, ethical review, and ongoing monitoring — never as a standalone verdict.

Evaluate constantly. Model evaluation starts when choosing a model for a use case, continues through development and testing — as you adjust inference parameters, engineer prompts, augment or customize with data, and add guardrails — and never stops in production. Combine **human reviewers with automated metrics**, and consider asking an LLM to critique its own outputs. Choose a standard set of metrics matched to your use case's priorities so you can validate progress on responsible AI across the lifecycle: measures target hallucinations, toxicity, and bias, with the

illegality measures split between **IP infringement** and **PII leakage**. The metric values you set as goals depend on each use case's requirements and what you know about the model and its training data.

TABLE 6.5 Content filters by risk

RISK	WHAT THE FILTER DOES
Hallucinations	Validate factual accuracy
Toxicity	Remove profanity and offensive language or images
Bias	Flag potentially biased language
Illegality	Mask or remove sensitive content
Misuse	Filter out potential prompt attacks or attempts at misconduct

Filters run on **both inputs and outputs** — removing unwanted content or flagging it for human follow-up, and blocking attempts to misuse the system on the way in. A practical note: given the breadth of tasks an FM supports, it is generally **less effort to control the output** of a generative model than to curate its training data and prompts.

Keeping humans in the loop. Human-in-the-loop (HITL) reviews integrate human oversight and decision-making into AI processes so that outputs stay accurate, ethical, and aligned with the intended use. Mistakes happen and language changes — but **LLMs are static**, and a human catches misalignment quickly. The greater the impact of getting it wrong, the more important the human review. HITL is critical during early proof-of-concept iterations and remains essential for regular reviews over time; once you know where humans fit, feedback can drive prompt changes, extra filtering, or automated steps where the LLM audits its own output — which is how HITL scales.

WORKED EXAMPLE – WHY THE HUMAN STAYS

Amazon search, 2019: customers searching *masks* wanted Halloween costumes. 2020: the same query needed to return surgical masks. No static model catches that shift on its own — a human in the loop does, fast. The lesson generalizes: monitor for the world changing under your model, not just for the model breaking.

WORKED EXAMPLE – THREE HITL PATTERNS IN PRODUCTION

Content moderation: the AI runs automated checks and flags potentially problematic content → human moderators make the final call (remove, approve, modify) → their decisions feed back to improve the AI's accuracy. **Translation QA:** the AI produces an initial translation → human translators correct for context, cultural nuance, and industry terminology → corrections improve the translation model. **Domain-specific validation (healthcare):** AI analyzes X-rays/MRIs or patient data and highlights concerns → medical professionals review and make the final diagnosis → their decisions refine the system. In every pattern the loop closes: human judgment both gates the output and trains the system.

TABLE 6.6 Monitoring, auditing, and feedback loops

MECHANISM	WHAT IT IS	EXAMPLE
Monitoring	Continuous, real-time observation and tracking of the system's performance, outputs, and behavior — dashboards and flags that alert the people managing the system	Real-time sentiment analysis flags negative interactions in a customer-service chat application
Auditing	Systematic, independent examination of the system, its processes, and outputs — verifying that what you saw in testing stays true in production	A quarterly audit examines responses across customer demographics to confirm equally helpful, respectful service
Feedback loops	Mechanisms feeding user input, performance data, and real-world outcomes back into the system — or pulling a human into the loop on a problem	After each interaction, users rate performance and flag potentially biased or inappropriate responses

6.4 How AWS generative AI services support responsible AI

AWS services ship with protections built in, and AWS invests in responsible AI across the foundation-model lifecycle — including collaboration with the global community and policymakers through the **White House voluntary AI commitments** and the **AI Safety Summit in the UK**, and support for **ISO 42001**, a new foundational standard for advancing responsible AI. On the practitioner side, the tooling splits between **Amazon SageMaker AI** and **Amazon Bedrock**.

TABLE 6.7 Key safeguarding tools in Amazon SageMaker AI

FEATURE	WHAT IT DOES	WHEN IT'S USED
SageMaker Model Cards	Standardized documentation for ML models — details, performance metrics, intended use cases; can disclose the distribution of sensitive attributes in training data to expose potential bias	Transparency, governance, compliance, reproducibility
SageMaker Clarify	Computes bias metrics and SHAP feature attributions; evaluates LLMs for toxicity and accuracy via automatic or human-based jobs	Explainability + evaluation, training through production
SageMaker Model Monitor	Via Clarify integration: continuous monitoring for **bias drift** and **feature attribution drift** in deployed models	Production monitoring
SageMaker Ground Truth	Human-based labeling of text, image, or video datasets — your own workforce or an Amazon/AWS Marketplace vendor	Development-time evaluation and fine-tuning data
Amazon Augmented AI (A2I)	Filters deployed-model predictions against your criteria and routes those that need it to a human workforce; results collected in Amazon S3	Production prediction review

SageMaker Clarify deserves the closest look. It provides a **model-agnostic feature attribution** approach using SHAP values to quantify each feature's contribution to the model's output. It detects bias types such as **demographic parity, equal opportunity, and disparate impact**, so biases can be found and mitigated in data or predictions. For LLMs it offers two evaluation modes: **automated evaluation jobs** covering text generation, text classification, question answering, and text summarization — with your own custom prompt dataset or a built-in one — to compare quality

and responsibility metrics for text-based FMs from **SageMaker JumpStart**; and **human-based evaluation jobs** that bring in your employees or industry subject-matter experts where human judgment is critical.

SAGEMAKER GROUND TRUTH	AMAZON A2I
development time ▪ Human-based labeling of text, image, or video datasets ▪ Workforce options: your own, Amazon, or an AWS Marketplace vendor ▪ Results evaluate model performance or become few-shot examples and fine-tuning datasets	*production time* ▪ Filters predictions from your deployed model before they reach users ▪ Your criteria (low confidence, random sampling) decide what goes to human review ▪ Reviewed responses land in an S3 bucket for the client application

COMMON PITFALL

The certification loves this distinction: **Amazon A2I reviews predictions on live, production applications**; **Ground Truth serves the iterative develop-and-evaluate cycle**. And for bias questions, the answer is **Clarify**, not Model Monitor — Model Monitor's own job is data drift and quality, gaining bias-drift detection only through Clarify integration.

Amazon Bedrock maps its features to the same safeguards. Its **choice of models** and side-by-side comparison let you pick the best fit — and see or change which model is in use. The **playgrounds** support experimenting with prompts, inference parameters, and models; **prompt management** turns the winners into organizational standards. The **evaluation feature** runs automatic or human-assisted jobs across common LLM use cases (text generation, classification, question answering, summarization) with metrics including **toxicity, accuracy, and robustness** — plus computed metrics that score how effectively a **knowledge base** returns relevant responses. Data curation runs through **Bedrock Knowledge Bases** (RAG over your own data) or model customization with data you know to be high quality.

TABLE 6.8 Amazon Bedrock Guardrails features

FILTER	WHAT IT DOES
Content-type filters	Filter categories of harmful user inputs and FM outputs; confidence-level sliders set breadth — set *hate* to low and content is filtered even when the model's confidence that it's hateful is low
Denied topics	Block output related to topics you specify — e.g. no medical advice
Word filters	Filter out specific words
Sensitive information filters	Mask or remove PII from generated outputs
Contextual grounding check	Helps verify the factual accuracy of output — the anti-hallucination check

Guardrails turn responsible-AI policy into enforcement: create **multiple guardrails tailored to different use cases** and apply them **across multiple foundation models**, giving a consistent user experience and standardized safety controls across generative AI applications. The module closes this section with a recorded demo of the Bedrock Guardrails console — creating and testing a guardrail.

WORKED EXAMPLE – THE SAMPLE EXAM QUESTION

A company wants to ensure their AI models are making fair and unbiased predictions across different demographic groups. Which Amazon SageMaker feature would be MOST effective? Key words: **fair and unbiased predictions**, **demographic groups**, **SageMaker feature**. Answer: **SageMaker Clarify** — purpose-built to detect bias in ML models and explain predictions, identifying and mitigating unfair bias across demographic groups. Why not the others: **Ground Truth** labels data, no bias detection; **A2I** runs human-review workflows, doesn't detect bias; **Model Monitor** catches data drift and quality issues in deployed models, not demographic bias specifically.

Module summary

- Responsible AI builds and uses systems that are people-centric, ethical, safe, and unbiased — with human oversight woven through the entire lifecycle, from data selection to production monitoring.
- AWS defines eight core dimensions in three groups: fairness, safety, privacy & security, and veracity & robustness protect people and data; transparency and explainability make AI open; controllability and governance actively manage it.
- Interpretability means understanding a model's internal mechanics (simple, transparent models); explainability gives a human-terms account of one output via model-agnostic techniques — and transparency always trades off against accuracy, privacy, safety, and security.
- Generative AI amplifies four challenges — hallucinations, toxicity, bias, and illegality (IP and PII in outputs). They can be managed, never eliminated; a use case with zero tolerance needs a different approach.
- Safeguards span the workflow: prompt engineering and data curation at the input, explainability and evaluation at the model, filtering and HITL at the output, monitoring/auditing/feedback loops in production — with data protection and governance across everything.
- The explainability toolkit: feature importance analysis, partial dependence plots, SHAP, LIME — or asking the LLM to justify its own answer.
- SageMaker AI tooling: Model Cards document models; Clarify explains and evaluates (bias, toxicity, accuracy — automatic or human jobs); Ground Truth labels during development; A2I sends production predictions to human review; Model Monitor + Clarify detect bias drift and feature attribution drift.
- Amazon Bedrock: model choice, playgrounds, and evaluation jobs (toxicity, accuracy, robustness; knowledge-base metrics), prompt management, Knowledge Bases for RAG — and Guardrails with content filters, denied topics, word filters, sensitive-information filters, and the contextual grounding check.

Review questions

1. A chatbot performs noticeably worse for one language community. Which responsible-AI dimension is at issue, and what practices address it?

Answer: Fairness — no harmful performance disparities across subpopulations (gender, ethnicity, age, language). Address it with diverse and representative training data, bias detection and mitigation, fairness metrics (demographic parity, predictive parity, CUAE — range 0–1), regular audits, and inclusive design with diverse stakeholders.

2. Distinguish transparency, interpretability, and explainability.

Answer: Transparency is the umbrella — the degree to which you can understand how the system works (documentation, data openness, UI disclosure, model cards, auditability). Interpretability is deep understanding of the internal mechanics — highly interpretable models are highly transparent. Explainability is a high-level, human-terms explanation of how a particular output was reached, typically via model-agnostic techniques on opaque-box models.

3. Why can't you remove bias from a pre-trained FM's training data, and what do you do instead?

Answer: You have no control over — or visibility into — a pre-trained FM's training data. Instead, evaluate the model for bias and fairness before adopting it, apply bias-mitigation techniques, monitor and audit outputs continuously, and keep humans reviewing. Data-curation techniques apply only where you customize a model or build your own.

4. A legal-brief generator invents case citations. Name the risk and four safeguards.

Answer: Hallucination — plausible but verifiably false output. Safeguards: fact-checking mechanisms, deterministic inference parameter settings, augmenting with relevant data (e.g. RAG over a verified citation database), and human-in-the-loop validation. Note the use-case rule: legal briefings have far less hallucination tolerance than creative storytelling.

5. Walk the generative AI application flow and name the safeguards at each point.

Answer: Prompt: prompt engineering, data augmentation, curation. FM: model transparency, explainability, evaluation. Output: content filtering, HITL reviews. Around the flow: monitoring, auditing, feedback loops — with data protection extending across the whole flow and responsible-AI guidelines, policies, and processes over everything.

6. What's the difference between SHAP and LIME?

Answer: SHAP calculates each feature's contribution to a prediction (Shapley values), explaining any ML model's output. LIME builds a simplified, interpretable version of the model around one specific prediction and explains that simpler model. Both are model-agnostic explainability techniques.

7. A deployed model should send its low-confidence predictions to humans before users see them. Which service, and how does the flow work?

Answer: Amazon A2I. The client application filters predictions with user-defined criteria (low confidence or random sampling); those needing review go to A2I's human workforce; results are collected in Amazon S3 for the client application. Ground Truth is the development-time counterpart — human labeling for evaluation and fine-tuning.

8. Which Amazon Bedrock Guardrails feature would you use to (a) block medical advice, (b) strip PII from outputs, (c) check outputs for factual accuracy?

Answer: (a) Denied topics, (b) sensitive information filters (mask or remove PII), (c) the contextual grounding check. Content-type filters with confidence sliders and word filters round out the set — and one guardrail can apply across multiple FMs.