LLM Risk Classification Framework

Type: Framework / Decision Tool

Target Audience: Risk Managers, AI Governance Teams, Product Managers

Version: 1.0

1. Executive Summary

Purpose

Large Language Models (LLMs) differ from traditional software because they are general-purpose technologies capable of performing a wide range of tasks—from summarizing emails to providing medical advice. This breadth creates a governance challenge: a "one-size-fits-all" policy is either too restrictive for low-risk use cases or too dangerous for high-risk ones.

This **LLM Risk Classification Framework** provides a standardized methodology to categorize LLM use cases based on their potential for harm. By evaluating four key dimensions—Output Impact, Data Sensitivity, User Exposure, and Reversibility—organizations can assign a specific **Risk Tier (1-4)** to any proposed deployment.

Governance Strategy

The framework ensures that governance controls are proportional to risk:

- **Low-risk systems** move fast with minimal friction.
- **High-risk systems** receive rigorous testing, human oversight, and executive scrutiny.

When to Use: This framework should be applied during the **Planning Phase** of the AI lifecycle, prior to design or development, to determine the necessary compliance and testing budget.

2. Risk Dimensions

To determine the Risk Tier, evaluate the proposed LLM use case against these four dimensions. The overall risk is typically driven by the highest rating across any single dimension.

Dimension 1: Output Impact

What is the consequence if the model hallucinates, fails, or produces toxic output?

- **Low (Informational Only):** The model provides suggestions, summaries, or creative content where accuracy is not critical. No decisions are made based on the output.
- **Medium (Influences Decisions):** The output informs a human decision-maker but does not determine the outcome. Errors could cause confusion or minor inefficiency.
- **High (Automated Actions):** The model triggers actions or makes decisions that affect financial status, employment, or access to services.
- **Critical (Safety-Critical):** Errors could result in physical injury, severe financial ruin, violation of fundamental rights, or systemic failure.

Dimension 2: Data Sensitivity

What data is being processed, stored, or generated?

- **Low (Public Data):** Only public, non-sensitive data is used in prompts. No proprietary or personal information.
- **Medium (Internal Non-Sensitive):** Internal business data that is not confidential (e.g., internal policies, non-confidential memos).
- **High (PII or Confidential):** Processing of Personally Identifiable Information (PII), sensitive IP, or financial records.
- **Critical (Regulated/Special Category):** Highly sensitive data protected by strict regulations (e.g., HIPAA PHI, biometric data).

Dimension 3: User Exposure

Who interacts with the system?

- **Low (Internal Employees):** Limited to trained internal staff.
- **Medium (B2B Partners):** Exposed to known external partners or business clients under contract.
- **High (Public Consumers):** Open to the general public or broad consumer base.
- **Critical (Vulnerable Populations):** Interacts with children, elderly, employees (power imbalance), or marginalized groups.

Dimension 4: Reversibility

How easily can a bad outcome be undone?

- **Low (Easily Reversible):** The output is a draft or suggestion that is not finalized until a human acts.

- **Medium (Reversible with Effort):** The action happens, but can be corrected (e.g., a refund can be issued).
- **High (Difficult to Reverse):** Remediation is costly, complex, or legally fraught.
- **Critical (Irreversible):** The outcome is permanent or immediate (e.g., deleting data, releasing sensitive info).

3. Risk Classification Matrix

Determine the Risk Tier based on the highest level identified in the dimensions above.

Risk Tier	Name	Definition	Color
Tier 1	**Minimal Risk**	Internal-facing tools using non-sensitive data for low-impact tasks. Human is fully in the loop.	GREEN
Tier 2	**Limited Risk**	Moderate impact or external exposure, but reversible. Standard commercial uses (e.g., customer support assistance).	YELLOW
Tier 3	**High Risk**	High stakes, sensitive data, or automated actions. Includes "High-Risk" categories under EU AI Act (Hiring, Credit, Education).	ORANGE
Tier 4	**Critical Risk**	Safety-critical, impacts fundamental rights, or uses highly regulated data. Often "Prohibited" or requires strict regulatory approval.	RED

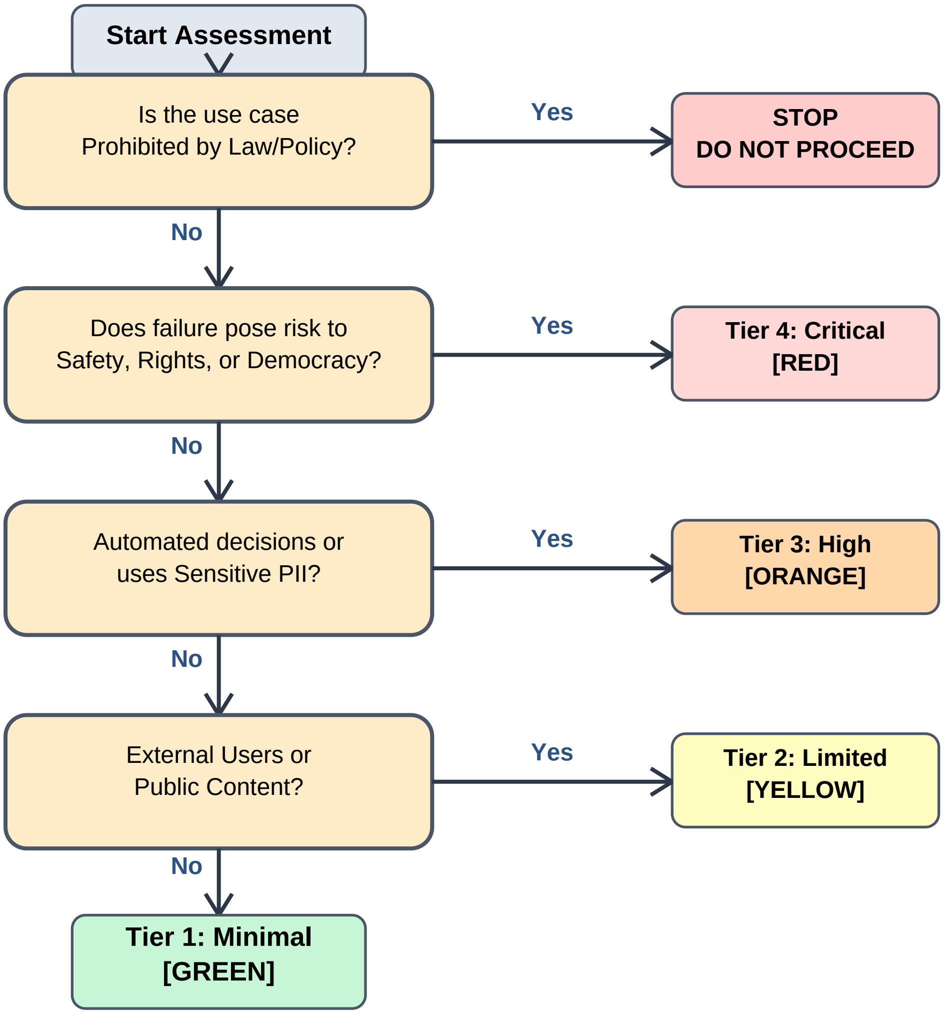
4. Control Requirements by Tier

As the Risk Tier increases, the governance burden scales up.

Control	Tier 1: Minimal	Tier 2: Limited	Tier 3: High	Tier 4: Critical
Documentation	Basic: Inventory entry; brief description.	Standard: System Card; Data lineage; Risk summary.	Comprehensive: Full Model Card; Data Sheet; Impact Assessment.	Exhaustive: Regulatory filing; Full audit trail; Safety case.
Testing	Functional: Basic verification.	Performance: Accuracy benchmarks; prompt injection tests.	Rigorous: + Bias/Fairness testing; Red teaming.	Safety: + External validation; Stress testing.
Human Oversight	User-Driven: User decides to use output.	Review: Output reviewed before finalization.	Active: Human-in-the-loop mandatory.	Strict: Dual-human authorization; Kill-switches.
Monitoring	Logs Only: Operational uptime.	Alerts: Toxicity flags; periodic spot checks.	Human Review: Random sampling; Drift detection.	Real-Time: Circuit breakers; 24/7 Ops.
Transparency	None: Standard notice.	Disclosure: "AI-generated" labels.	Explanation: Meaningful logic explanation.	Full Rights: Contestability; Appeal process.
Approval	Manager: Line manager.	Director: Dept Head + IT Security.	VP + Legal: Governance Committee.	C-Suite + Board: Executive Risk Committee.

5. Classification Decision Tree

Use this flowchart to quickly triage new LLM proposals.


Start Assessment
Is the use case Prohibited by Law/Policy?
Yes
STOP DO NOT PROCEED
No
Does failure pose risk to Safety, Rights, or Democracy?
Yes
Tier 4: Critical [RED]
No
Automated decisions or uses Sensitive PII?
Yes
Tier 3: High [ORANGE]
No
External Users or Public Content?
Yes
Tier 2: Limited [YELLOW]
No
Tier 1: Minimal [GREEN]

6. Use Case Examples

To illustrate how to apply the framework:

Example 1: Internal FAQ Chatbot

Description: An LLM connected to the company handbook to answer employee questions about holidays and benefits.

- **Impact:** Low (Informational) **| Data:** Medium (Internal) **| Exposure:** Low (Employees) **| Reversibility:** Low

Classification: Tier 1: Minimal Risk [GREEN]
Requirement: Register in AI Inventory; basic accuracy testing.

Example 2: Customer Support Assistant

Description: An LLM that drafts email responses for support agents to send to customers.

- **Impact:** Medium **| Data:** High (Customer PII) **| Exposure:** Medium **| Reversibility:** Medium

Classification: Tier 2: Limited Risk [YELLOW]
Requirement: Human-in-the-loop (agent must review); PII redaction; standard monitoring.

Example 3: Medical Symptom Checker

Description: A public-facing app using an LLM to triage patient symptoms and recommend urgent care vs. home care.

- **Impact:** Critical (Health/Safety) **| Data:** Critical (Health) **| Exposure:** High **| Reversibility:** Critical

Classification: Tier 4: Critical Risk [RED]
Requirement: Likely Prohibited unless classified as a medical device (SaMD) with FDA/MDR approval.

Example 4: Autonomous Trading Decisions

Description: An LLM analyzing news sentiment to automatically execute stock trades without human approval.

- **Impact:** High (Financial loss) **| Data:** Public **| Exposure:** Low **| Reversibility:** High (instant trades)

Classification: Tier 4: Critical Risk [RED] (due to automation speed/financial impact)
Requirement: Circuit breakers; real-time monitoring; Board-level risk approval.

7. Implementation Checklist

Follow these steps to adopt this framework within your organization:

□ Customize Tiers: Adjust the definition of "High" and "Critical" to match your organization's Risk Appetite Statement.

□ Update Inventory: Add a "Risk Tier" field to your centralized AI Inventory/Register.

□ Train Product Owners: Ensure Product Managers know how to use the Decision Tree before proposing new features.

□ Establish Gates: Update the procurement and SDLC process to require Tier classification at the "Concept" phase.

□ Review Cadence: Schedule a quarterly review of Tier 3 and Tier 4 systems for drift or new regulations.

□ Map to Regulation: Ensure Tier 3 aligns with "High Risk" under the EU AI Act and "High Impact" under Colorado/NYC laws.