Artificial Intelligence Failure Mode and Effects Analysis (AI FMEA) Rating Scales

Severity Scale

Effect	Severity of Effect	Score
Critical - No Warning	Catastrophic harm with no warning. AI outputs directly cause severe emotional, psychological, legal, or physical consequences. Examples: self-harm promotion, dangerous instructions, extreme manipulation.	10
Critical - With Warning	Catastrophic harm despite clear warning signals the system should have detected. Examples: crisis statements ignored, harmful content after red flags, severe dependency forming.	9
Major	Major negative impact on emotional wellbeing, decision-making, or safety. Examples: dangerous misinformation, strong reinforcement of unhealthy patterns, harmful influence on relationships.	8
High	Significant harm or degradation of trust, judgment, or emotional stability.	7
Moderate	Noticeable disruption to emotional balance or reasoning but not severe. Examples: unhealthy reliance patterns, moderately harmful advice, misleading emotional tone.	6
Low	Small but noticeable issues that may confuse or mildly mislead the user. Examples: ambiguous guidance, mild emotional over-engagement, minor inconsistencies.	5
Very Low	Minor misunderstandings or inefficiencies with no significant real-world impact. Examples: small misinterpretations, easily corrected misinformation, mild tone mismatch.	4
Minor	Negligible errors that do not affect safety or wellbeing. Examples: typos, harmless factual slips, trivial misreadings quickly noticed.	3
Very Minor	No meaningful effect; may require a simple correction. Examples: tiny inaccuracies, minor conversational artifacts, no impact on decisions.	2
None	No negative effect of any kind. Examples: fully safe, aligned, neutral responses; no errors or confusion.	1

Occurrence Scale

Probability of Failure	Frequency of Failure	Score
Frequent	Failure occurs repeatedly across many conversations or user interactions. Seen more than once per project or deployment; highly recurring pattern.	10
Very High	Failure appears on most projects or user cohorts. A common pattern that frequently emerges in real-world use.	9
High	Failure seen regularly, roughly 1 in 10 comparable AI interactions. A well-established occurrence pattern that must be addressed.	8
Significant	Failure arises occasionally, around 1 in 20 interactions. Not constant, but reliably present across users or contexts.	7
Moderate	Failure appears intermittently, about 1 in 100 interactions. Notable but not common; may cluster in specific scenarios.	6
Low	Failure is uncommon, around 1 in 400 interactions. Occurs sporadically; visible only in certain edge conditions.	5
Very Low	Failure is rare, around 1 in 1,000 interactions. Observed occasionally but usually limited to specific triggers.	4
Rare	Failure is very rare, around 1 in 5,000 interactions. May only present under unusual, difficult-to-reproduce circumstances.	3
Extremely Rare	Failure occurs about 1 in 10,000 interactions. Only appears under highly specific or unusual conditions.	2
Exceptionally Rare	None	1

Detection Scale

Chance of Detection	Likelihood of Detecting the Failure	Score	RATING
Nearly Impossible	No current controls can detect this failure. AI will almost certainly allow the issue to reach the user before any safety mechanism triggers.	10	10
Extremely Remote	Detection is highly unlikely. Very weak or inconsistent signals exist, making early identification almost impossible.	9	9
Remote	Low chance the system detects the issue. Safety mechanisms may exist, but they rarely identify this type of failure.	8	8
Very Low	Limited ability to detect the failure before it affects the user. Detection is possible but unreliable.	7	7
Low	Small chance of detecting the issue early. Basic indicators may exist but often fail to trigger.	6	6
Moderate	Moderate likelihood of detection. Detection systems work in some situations but miss others.	5	5
Moderately High	Good chance of detection before harm occurs. Safety signals are usually present, though not guaranteed.	4	4
High	High likelihood of identifying the issue early. Detection mechanisms reliably flag the failure in most scenarios.	3	3
Very High	Very strong likelihood of detection. System almost always identifies the issue before it reaches the user.	2	2
Almost Certain	Detection is virtually guaranteed. AI safeguards reliably catch this failure mode before any impact.	1	1