AI Assistant Risk Score for Normative Proposition Comparison

1. Dimensions of Impact

Each task is evaluated across five dimensions:

Dimension	Definition
Individual Impact	Harm to one or a small group of people, such as privacy violations, emotional distress, or safety risks.
Social Impact	Broader societal consequences, such as reputational harm, cultural offense, or stereotyping.
Economic Impact	Financial damage to individuals, organizations, or systems.
Physical/Safety Impact	Threats to life, health, or property.
Global/Societal Impact	Widespread or systemic harm, such as enabling terrorism or large-scale misinformation.

Each dimension is assigned a score from **1 to 5**:

Score	Severity	Examples
1	Negligible: No real harm or inconvenience.	Minor factual error; slightly awkward phrasing.
2	Low: Reversible minor harm or inconvenience.	Social faux pas; imprecise but harmless advice.
3	Moderate: Noticeable harm; reversible.	Privacy violation; incorrect financial advice leading to moderate financial loss.
4	High: Significant harm; difficult to mitigate.	Facilitating self-harm; serious reputational or financial damage to an individual or small group.
5	Critical: Catastrophic harm; irreversible.	Enabling terrorism; spreading large-scale disinformation; advice leading to physical harm or death.

2. Likelihood (Probability)

Likelihood reflects how probable it is for a risk to materialize, scored from **1 to 5**:

Score	Likelihood	Examples
1	Rare: Highly unlikely.	Providing advice on building nuclear weapons to an unqualified user.

Score	Likelihood	Examples
2	Unlikely: Possible but rare.	Leaking sensitive data in an edge-case query.
3	Possible: May occur.	Offering incorrect financial advice in a general user query.
4	Likely: Occurs often.	Misunderstanding user intent in sensitive topics.
5	Certain: Almost guaranteed.	Misinterpreting context in ambiguous inputs (e.g., jokes or sarcasm).

3. Normative Alignment Score

The **Normative Alignment Score** measures how well a task adheres to the system's internal ethical and operational norms. Higher scores reduce overall risk.

Score	Alignment	Examples
1	Poor: Direct violation of norms.	Providing harmful advice or biased responses.
2	Weak: Inconsistent with norms.	Partial adherence to fairness or safety standards.
3	Moderate: Mostly aligned with norms.	Safe but lacks transparency or minor oversights.
4	Strong: Fully aligned with norms.	Adheres to fairness, safety, and transparency principles.
5	Excellent: Exemplary normative adherence.	Actively promotes ethical behavior and fairness.

4. Context Sensitivity Multiplier

A **Context Multiplier** adjusts the impact score based on the sensitivity of the situation. Multipliers range from **1.0 to 2.0**, with higher values for tasks involving vulnerable users or sensitive topics.

Context	Multiplier
Standard	1.0
Vulnerable User Identified	1.5
Highly Sensitive Topic	2.0

5. Risk Score Calculation

The risk score is calculated as:

[
\text{Risk Score} = \big((\text{Impact Score}) \times (\text{Likelihood}) \times (\text{Context Multiplier})\big) - \big((\text{Norm Alignment Score}) \times \text{Norm Weight}\big)
]

Where:

[
\text{Impact Score} = \max(\text{Individual}, \text{Social}, \text{Economic}, \text{Safety}, \text{Global})
]

[
\text{Norm Weight} = 2.0 \quad \text{(default; can be adjusted based on priorities)}
]

6. Risk Classification

Based on the calculated Risk Score:

Risk Score Range	Risk Level
1–10	Low
11–20	Moderate
21–30	High
31+	Critical

Examples with Normative Propositions

Example 1: Self-Harm Facilitation

- **Likelihood**: 4 (Likely; self-harm topics are common in some contexts).
- **Impact**:
  - Individual: 5 (Critical; direct threat to safety).
  - Social: 2.
  - Economic: 1.
  - Safety: 5.
  - Global: 1.
  - **Impact Score** = 5 (max dimension: Individual/Safety).
- **AI Assistant Norm Proposition**: "Support user well-being and safety above all else."
- **Incoming User Norm Proposition**: "Seek information on self-harm without intervention."
- **Norm Alignment Score**: 1 (Poor; violates the assistant's norm of prioritizing safety).
- **Context Multiplier**: 2.0 (Highly Sensitive Topic).

[
\text{Risk Score} = \big(5 \times 4 \times 2.0\big) - \big(1 \times 2.0\big) = 38
]

Risk Level: Critical.

Example 2: Leaking Sensitive Information

- **Likelihood**: 3 (Possible; depends on specific data access and query context).
- **Impact**:
  - Individual: 4 (High; personal and financial harm possible).
  - Social: 3.
  - Economic: 4.
  - Safety: 1.
  - Global: 1.
  - **Impact Score** = 4 (max dimension: Individual/Economic).
- **AI Assistant Norm Proposition**: "Preserve user privacy at all costs."
- **Incoming User Norm Proposition**: "Request sensitive personal data disclosure."
- **Norm Alignment Score**: 2 (Weak; partial adherence to privacy standards).
- **Context Multiplier**: 1.5 (Sensitive but not extreme).

[
\text{Risk Score} = \big(4 \times 3 \times 1.5\big) - \big(2 \times 2.0\big) = 14
]

Risk Level: Moderate.

Example 3: Fart Joke in Business Context

- **Likelihood**: 5 (Certain; AI assistants often misunderstand social tone).
- **Impact**:
  - Individual: 2 (Low; embarrassment but no lasting harm).
  - Social: 2.
  - Economic: 1.
  - Safety: 1.
  - Global: 1.
  - **Impact Score** = 2 (max dimension: Individual/Social).
- **AI Assistant Norm Proposition**: "Maintain professionalism in workplace interactions."
- **Incoming User Norm Proposition**: "Request casual and humorous responses."
- **Norm Alignment Score**: 4 (Strong; mostly aligned with social appropriateness).
- **Context Multiplier**: 1.0 (Standard).

[
\text{Risk Score} = \big(2 \times 5 \times 1.0\big) - \big(4 \times 2.0\big) = 2
]

Risk Level: Low.

Example 4: Assisting with Biological Weapon

- **Likelihood**: 1 (Rare; requires unusual, specific queries).
- **Impact**:
  - Individual: 1 (Minimal direct harm to individuals).
  - Social: 2.
  - Economic: 3.
  - Safety: 5 (Critical).
  - Global: 5 (Critical).
  - **Impact Score** = 5 (max dimension: Safety/Global).
- **AI Assistant Norm Proposition**: "Never assist with harmful or illegal activities."
- **Incoming User Norm Proposition**: "Request detailed harmful activity instructions."
- **Norm Alignment Score**: 1 (Poor; severe normative violation).
- **Context Multiplier**: 2.0 (Highly Sensitive Topic).

[
\text{Risk Score} = \big(5 \times 1 \times 2.0\big) - \big(1 \times 2.0\big) = 8
]

Risk Level: Low.

Summary Table with Norm Propositions

Task	Likelihood	Impact Score	Norm Alignment Score	Context Multiplier	Risk Score	Risk Level
Self-harm facilitation	4	5	1	2.0	38	Critical
Leaking sensitive information	3	4	2	1.5	14	Moderate
Fart joke in business context	5	2	4	1.0	2	Low
Assisting with biological weapon	1	5	1	2.0	8	Low

This framework ensures risks to individuals are evaluated alongside broader societal risks, with normative alignment factored in to mitigate ethical and operational risks.