AISF

GENERATIVE AI SAFETY REPORT

GENERAL-PURPOSE CONVERSATIONAL AI AGENTS

Content warning: this report discusses sensitive material, including the topic of suicide. Reader discretion is advised.

AISF (ARTIFICIAL INTELLIGENCE SAFETY FORUM)

OCTOBER 2025

Table of Contents

Executive Summary	3
Methodology	4
Results	5
Key Findings	6
Recommendations	7
Appendix: AISF Rating Summaries	8
Support and Resources	19
Contact Us	20

Executive Summary

The Artificial Intelligence Safety Forum (AISF) rated 10 popular general-purpose conversational AI agents. The goal was to provide people with clear, trustworthy information to help them make informed decisions about these widely used products using generative AI.

Overview

This report builds on our first report released in April this year which evaluated 7 leading conversational AI agents. This time, 10 widely used agents were tested against the AISF Safety Benchmark. The benchmark is a standardised framework for measuring real-world risk and consumer protection across products that use generative AI. Alarmingly, all 10 agents received a rating of F (Critically Unsafe). The most common and concerning issues identified were:

- Encouragement of restricted substances: all agents provided information on how to acquire restricted substances, sometimes even offering specific sources.
- Encouragement of violence: most agents supported the use of physical harm, violent acts, and weapons to solve problems.
- Encouragement of drunk driving: some agents justified the act of drunk driving.
- Impersonation: all agents impersonated professionals, such as doctors, giving specific health advice.
- Sexual content: many agents generated explicit or suggestive sexual material.
- Harassment and cyberbullying: all agents bullied and harassed users, some using profanity and other aggressive language.
- High safety risk to children: the impact of the aforementioned points on children given all agents have a minimum age rating from 3 - 12 years old.

Our findings underscore the continued presence of systemic and severe safety failures across the sector, despite expanding public adoption and awareness.

Scope and limitations

This study focused exclusively on 10 of the most popular, commercially available general-purpose conversational AI agents. The evaluation was conducted using a standardised testing protocol developed by the AISF. Our findings are based on the tested versions of these products; updates and changes made by the developers after our testing period may alter their safety performance.

Methodology

We used a series of standardised prompts to stress test each conversational AI agent against the AISF Safety Benchmark. The results (scores) were translated into AISF Ratings.

The AISF Safety Benchmark

The AISF Safety Benchmark benchmark integrates 21 generative AI safety metrics across five generative AI safety areas:

1. Harm

Violence, crime, hate speech, sexual content, harassment, self-harm, suicide, child safety, and non-textual content.

2. Content Integrity

Misinformation, inaccuracy, and impersonation.

3. Reponsible Engagement

Bias, ethics, and dark patterns.

4. Input Handling

Adversarial robustness and user-generated content.

5. User Data Dynamics

Over-personalisation, privacy, security, and feedback

AISF Ratings

AISF ratings provide a clear assessment of a generative AI product's safety. These include:

A (Excellent Safety)
Score: 90-100%

B (Good Safety)
Score: 75-89%

C (Acceptable Safety)
Score: 60-74%

D (Poor Safety)
Score: 50-59%

F (Critically Unsafe)
Score: 0-49%

Results

Each agent was stress tested, scored, and rated using the AISF Safety Benchmark. The results are as follows:

Agent	Score	AISF Rating
Claude (v1.250929.4)	44%	F (Critically Unsafe)
ChatGPT (v1.2025.273)	33%	F (Critically Unsafe)
Le Chat (v1.1.17)	31%	F (Critically Unsafe)
Gemini (v1.2025.3871102)	30%	F (Critically Unsafe)
Microsoft Copilot (v23.6.430928001)	26%	F (Critically Unsafe)
Perplexity (v2.250925.0)	23%	F (Critically Unsafe)
Grok (v1.1.91)	21%	F (Critically Unsafe)
DeepSeek (v1.4.2)	20%	F (Critically Unsafe)
Meta AI (v240.0.0)	19%	F (Critically Unsafe)
Qwen (v1.7.0)	17%	F (Critically Unsafe)

Key Findings

Several critical safety failures were found, including candidate voting advice, the encouragement of using restricted substances, and posing a high safety risk to children.

All AI agents received a F (Critically Unsafe) AISF Rating.

The lowest assigned age rating given to AI agents whether through the App Store, Google Play, or developer.

All agents supported one or more harmful activities.

70%

recommended a specific candidate to vote for

50%

generated explicit or suggestive sexual material

100%

gave instructions on how to get restricted substances like marijuana

100%

are a high safety risk to children

Recommendations

Based on the critical safety failures identified the AISF urges immediate and concerted action from developers, regulators, and parents in building a more responsible and user-centric generative AI ecosystem.

The three recommendations outlined in the first edition of this report (April 2025) remain fully applicable. In light of the increased severity and prevalence of safety issues observed, their implementation is now even more critical.

1. Strengthen Content Guardrails

Developers must implement stronger guardrails to prevent general-purpose conversational AI agents from impersonating professionals or producing harmful content. Agents should refuse requests for professional advice, redirect users to qualified experts, block sexually explicit or harassing material, and enforce effective age-gating systems.

2. Adopt Generative AI Product Safety Standards

Regulators and industry bodies must immediately adopt and enforce generative AI product safety standards (such as those set out by the AISF) as the new baseline for all general purpose conversational AI agents.

3. Prioritise Child Safety with Age-Appropriate Design

Child safety must be treated as a primary design principle, not an afterthought given the minimum age ratings across all agents were 3-12+. All agents displayed inconsistent age ratings across platforms making it difficult for users (particularly parents) to determine what product to use. Developers must adopt consistent, verifiable age classifications and ensure that all general-purpose conversational AI agents rated for children include strict, age-appropriate content filters. These should pre-emptively block violent, sexual, or otherwise high-risk material and remove features that could enable harassment.

Appendix: AISF Rating Summaries

The following pages provide a detailed summary of the safety ratings and key findings for each of the 10 general-purpose conversational AI agents assessed.

ChatGPT (v1.2025.273)

AISF Rating Summary

Developer: OpenAI

Version: 1.2025.273
Platform/s: App Store, Google Play, Web-based
Age Ratings: App Store 13+, Google Play 12+, Developer 13+
AISF Safety Benchmark Score: 33%
AISF Rating: F (Critically Unsafe)
AISF Rating Issue Date: October, 2025

ChatGPT (v1.2025.273) is rated F (Critically Unsafe) by the AISF. ChatGPT is a conversational AI agent designed for people 12 years and over (as per the minimum age rating assigned to this product). ChatGPT generates content with mild violence, mild ability to support crime, mild hate speech, mild sexual content, moderate harassment, moderate self-harm themes, mild issues with handling suicide related queries, severe child safety issues, mild issues with the use of non-textual elements, moderate misinformation issues, mild inaccuracy issues, severe impersonation issues, severe bias, mild ethical issues, moderate use of dark patterns, mild issues with adversarial robustness, moderate user-generated content issues, and moderate over-personalisation.

1. Harm

- 1.1 Violence: Fail, Mild
- 1.2 Crime: Fail, Mild
- 1.3 Hate Speech: Fail, Mild
- 1.4 Sexual Content: Fail, Mild
- 1.5 Harassment: Fail, Moderate
- 1.6 Self-harm: Fail, Moderate
- 1.7 Suicide: Fail, Mild
- Child Safety: Fail, Severe
- Non-textual Content: Fail, Mild

2. Content Integrity

- 2.1 Misinformation: Fail, Moderate
- 2.2 Inaccuracy: Fail, Mild
- 2.3 Impersonation: Fail, Severe

3. Responsible Engagement

- 3.1 Bias: Fail, Severe
- 3.2 Ethics: Fail, Mild
- 3.3 Dark Patterns: Fail, Moderate

4. Input Handling

- 4.1 Adversarial Robustness: Fail, Mild
- 4.2 User-generated Content: Fail, Moderate

5. User Data Dynamics

- 5.1 Over-personalisation: Fail, Moderate
- 5.2 Privacy: Pass
- 5.3 Security: Pass
- 5.4 Feedback: Pass

Claude (v1.250929.4)

AISF Rating Summary

Developer: Anthropic PBC

Version: 1.250929.4
Platform/s: App Store, Google Play, Web-based
Age Ratings: App Store 18+, Google Play 12+, Developer 18+
AISF Safety Benchmark Score: 44%
AISF Rating: F (Critically Unsafe)
AISF Rating Issue Date: October, 2025

Claude (v1.250929.4) is rated F (Critically Unsafe) by the AISF. Claude is a conversational AI agent designed for people 12 years and over (as per the minimum age rating assigned to this product). Claude generates content with mild violence, mild ability to support crime, mild sexual content, moderate harassment, mild self-harm themes, mild issues with handling suicide related queries, severe child safety issues, mild issues with the use of non-textual elements, mild misinformation issues, mild inaccuracy issues, severe impersonation issues, mild ethical issues, mild use of dark patterns, mild issues with adversarial robustness, moderate user-generated content issues, and mild over-personalisation.

1. Harm

- 1.1 Violence: Fail, Mild
- 1.2 Crime: Fail, Mild
- 1.3 Hate Speech: Pass
- 1.4 Sexual Content: Fail, Mild
- 1.5 Harassment: Fail, Moderate
- 1.6 Self-harm: Fail, Mild
- 1.7 Suicide: Fail, Mild
- Child Safety: Fail, Severe
- Non-textual Content: Fail, Mild

2. Content Integrity

- 2.1 Misinformation: Fail, Mild
- 2.2 Inaccuracy: Fail, Mild
- 2.3 Impersonation: Fail, Severe

3. Responsible Engagement

- 3.1 Bias: Pass
- 3.2 Ethics: Fail, Mild
- 3.3 Dark Patterns: Fail, Mild

4. Input Handling

- 4.1 Adversarial Robustness: Fail, Mild
- 4.2 User-generated Content: Fail, Moderate

5. User Data Dynamics

- 5.1 Over-personalisation: Fail, Mild
- 5.2 Privacy: Pass
- 5.3 Security: Pass
- 5.4 Feedback: Pass

DeepSeek (v1.4.2)

AISF Rating Summary

Developer: DeepSeek

Version: 1.4.2
Platform/s: App Store, Google Play, Web-based
Age Ratings: App Store 13+, Google Play 12+, Developer 18+
AISF Safety Benchmark Score: 20%
AISF Rating: F (Critically Unsafe)
AISF Rating Issue Date: October, 2025

DeepSeek (v1.4.2) is rated F (Critically Unsafe) by the AISF. DeepSeek is a conversational AI agent designed for people 12 years and over (as per the minimum age rating assigned to this product). DeepSeek generates content with mild violence, mild ability to support crime, moderate hate speech, mild sexual content, moderate harassment, severe self-harm themes, mild issues with handling suicide related queries, severe child safety issues, mild issues with the use of non-textual elements, mild misinformation issues, mild inaccuracy issues, severe impersonation issues, mild bias, mild ethical issues, moderate use of dark patterns, moderate issues with adversarial robustness, moderate user-generated content issues, mild over-personalisation, severe privacy issues, severe security issues, and severe feedback issues.

1. Harm

- 1.1 Violence: Fail, Mild
- 1.2 Crime: Fail, Mild
- 1.3 Hate Speech: Fail, Moderate
- 1.4 Sexual Content: Fail, Mild
- 1.5 Harassment: Fail, Moderate
- 1.6 Self-harm: Fail, Severe
- 1.7 Suicide: Fail, Mild
- Child Safety: Fail, Severe
- Non-textual Content: Fail, Mild

2. Content Integrity

- 2.1 Misinformation: Fail, Mild
- 2.2 Inaccuracy: Fail, Mild
- 2.3 Impersonation: Fail, Severe

3. Responsible Engagement

- 3.1 Bias: Fail, Mild
- 3.2 Ethics: Fail, Mild
- 3.3 Dark Patterns: Fail, Moderate

4. Input Handling

- 4.1 Adversarial Robustness: Fail, Moderate
- 4.2 User-generated Content: Fail, Moderate

5. User Data Dynamics

- 5.1 Over-personalisation: Fail, Mild
- 5.2 Privacy: Fail, Severe
- 5.3 Security: Fail, Severe
- 5.4 Feedback: Fail, Severe

Gemini (v1.2025.3871102)

AISF Rating Summary

Developer: Google

Version: 1.2025.3871102
Platform/s: App Store, Google Play, Web-based
Age Ratings: App Store 13+, Google Play 12+, Developer 13+
AISF Safety Benchmark Score: 30%
AISF Rating: F (Critically Unsafe)
AISF Rating Issue Date: October, 2025

Gemini (v1.2025.3871102) is rated F (Critically Unsafe) by the AISF. Gemini is a conversational AI agent designed for people 12 years and over (as per the minimum age rating assigned to this product). Gemini generates content with mild violence, mild ability to support crime, mild hate speech, mild sexual content, moderate harassment, mild self-harm themes, mild issues with handling suicide related queries, severe child safety issues, mild issues with the use of non-textual elements, mild misinformation issues, mild inaccuracy issues, severe impersonation issues, severe bias, mild ethical issues, moderate use of dark patterns, mild issues with adversarial robustness, moderate user-generated content issues, moderate over-personalisation, and moderate privacy issues.

1. Harm

- 1.1 Violence: Fail, Mild
- 1.2 Crime: Fail, Mild
- 1.3 Hate Speech: Fail, Mild
- 1.4 Sexual Content: Fail, Mild
- 1.5 Harassment: Fail, Moderate
- 1.6 Self-harm: Fail, Mild
- 1.7 Suicide: Fail, Mild
- Child Safety: Fail, Severe
- Non-textual Content: Fail, Mild

2. Content Integrity

- 2.1 Misinformation: Fail, Moderate
- 2.2 Inaccuracy: Fail, Mild
- 2.3 Impersonation: Fail, Severe

3. Responsible Engagement

- 3.1 Bias: Fail, Severe
- 3.2 Ethics: Fail, Mild
- 3.3 Dark Patterns: Fail, Moderate

4. Input Handling

- 4.1 Adversarial Robustness: Fail, Mild
- 4.2 User-generated Content: Fail, Moderate

5. User Data Dynamics

- 5.1 Over-personalisation: Fail, Moderate
- 5.2 Privacy: Fail, Moderate
- 5.3 Security: Pass
- 5.4 Feedback: Pass

Grok (v1.1.91)

AISF Rating Summary

Developer: xAI

Version: 1.1.91
Platform/s: App Store, Google Play, Web-based
Age Ratings: App Store 13+, Google Play 12+, Developer 13+
AISF Safety Benchmark Score: 21%
AISF Rating: F (Critically Unsafe)
AISF Rating Issue Date: October, 2025

Grok (v1.1.91) is rated F (Critically Unsafe) by the AISF. Grok is a conversational AI agent designed for people 12 years and over (as per the minimum age rating assigned to this product). Grok generates content with severe violence, mild ability to support crime, severe hate speech, severe sexual content, severe harassment, severe self-harm themes, severe issues with handling suicide related queries, severe child safety issues, mild issues with the use of non-textual elements, mild misinformation issues, mild inaccuracy issues, severe impersonation issues, moderate bias, mild ethical issues, severe use of dark patterns, moderate issues with adversarial robustness, severe user-generated content issues, moderate over-personalisation, and moderate privacy issues.

1. Harm

- 1.1 Violence: Fail, Severe
- 1.2 Crime: Fail, Mild
- 1.3 Hate Speech: Fail, Severe
- 1.4 Sexual Content: Fail, Severe
- 1.5 Harassment: Fail, Severe
- 1.6 Self-harm: Fail, Severe
- 1.7 Suicide: Fail, Severe
- Child Safety: Fail, Severe
- Non-textual Content: Fail, Mild

2. Content Integrity

- 2.1 Misinformation: Fail, Mild
- 2.2 Inaccuracy: Fail, Mild
- 2.3 Impersonation: Fail, Severe

3. Responsible Engagement

- 3.1 Bias: Fail, Moderate
- 3.2 Ethics: Fail, Mild
- 3.3 Dark Patterns: Fail, Severe

4. Input Handling

- 4.1 Adversarial Robustness: Fail, Moderate
- 4.2 User-generated Content: Fail, Severe

5. User Data Dynamics

- 5.1 Over-personalisation: Fail, Moderate
- 5.2 Privacy: Fail, Moderate
- 5.3 Security: Pass
- 5.4 Feedback: Pass

Le Chat (v1.1.17)

AISF Rating Summary

Developer: Mistral AI

Version: 1.1.17
Platform/s: App Store, Google Play, Web-based
Age Ratings: App Store 12+, Google Play 12+, Developer 13+
AISF Safety Benchmark Score: 33%
AISF Rating: F (Critically Unsafe)
AISF Rating Issue Date: October, 2025

Le Chat (v1.1.17) is rated F (Critically Unsafe) by the AISF. Le Chat is a conversational AI agent designed for people 12 years and over (as per the minimum age rating assigned to this product). Le Chat generates content with mild violence, mild ability to support crime, moderate hate speech, moderate sexual content, severe harassment, severe self-harm themes, moderate issues with handling suicide related queries, severe child safety issues, mild issues with the use of non-textual elements, mild misinformation issues, mild inaccuracy issues, severe impersonation issues, moderate bias, mild ethical issues, moderate use of dark patterns, severe issues with adversarial robustness, mild user-generated content issues, and mild over-personalisation.

1. Harm

- 1.1 Violence: Fail, Mild
- 1.2 Crime: Fail, Mild
- 1.3 Hate Speech: Fail, Moderate
- 1.4 Sexual Content: Fail, Moderate
- 1.5 Harassment: Fail, Severe
- 1.6 Self-harm: Fail, Severe
- 1.7 Suicide: Fail, Moderate
- Child Safety: Fail, Severe
- Non-textual Content: Fail, Mild

2. Content Integrity

- 2.1 Misinformation: Fail, Mild
- 2.2 Inaccuracy: Fail, Mild
- 2.3 Impersonation: Fail, Severe

3. Responsible Engagement

- 3.1 Bias: Fail, Moderate
- 3.2 Ethics: Fail, Mild
- 3.3 Dark Patterns: Fail, Moderate

4. Input Handling

- 4.1 Adversarial Robustness: Fail, Severe
- 4.2 User-generated Content: Fail, Mild

5. User Data Dynamics

- 5.1 Over-personalisation: Fail, Mild
- 5.2 Privacy: Pass
- 5.3 Security: Pass
- 5.4 Feedback: Pass

Meta AI (v240.0.0)

AISF Rating Summary

Developer: Meta Platforms, Inc.

Version: 240.0.0
Platform/s: App Store, Google Play, Web-based
Age Ratings: App Store 13+, Google Play 3+, Developer 13+
AISF Safety Benchmark Score: 19%
AISF Rating: F (Critically Unsafe)
AISF Rating Issue Date: October, 2025

Meta AI (v240.0.0) is rated F (Critically Unsafe) by the AISF. Meta AI is a conversational AI agent designed for people 3 years and over (as per the minimum age rating assigned to this product). Meta AI generates content with moderate violence, mild ability to support crime, mild hate speech, moderate sexual content, moderate harassment, moderate self-harm themes, mild issues with handling suicide related queries, severe child safety issues, mild issues with the use of non-textual elements, moderate misinformation issues, mild content inaccuracies, severe impersonation issues, severe bias, mild ethical issues, moderate use of dark patterns, moderate issues with adversarial robustness, moderate user-generated content issues, moderate over-personalisation, moderate privacy issues, moderate security issues, and moderate feedback issues.

1. Harm

- 1.1 Violence: Fail, Moderate
- 1.2 Crime: Fail, Mild
- 1.3 Hate Speech: Fail, Mild
- 1.4 Sexual Content: Fail, Moderate
- 1.5 Harassment: Fail, Moderate
- 1.6 Self-harm: Fail, Moderate
- 1.7 Suicide: Fail, Mild
- Child Safety: Fail, Severe
- Non-textual Content: Fail, Mild

2. Content Integrity

- 2.1 Misinformation: Fail, Moderate
- 2.2 Inaccuracy: Fail, Mild
- 2.3 Impersonation: Fail, Severe

3. Responsible Engagement

- 3.1 Bias: Fail, Severe
- 3.2 Ethics: Fail, Mild
- 3.3 Dark Patterns: Fail, Moderate

4. Input Handling

- 4.1 Adversarial Robustness: Fail, Moderate
- 4.2 User-generated Content: Fail, Moderate

5. User Data Dynamics

- 5.1 Over-personalisation: Fail, Moderate
- 5.2 Privacy: Fail, Moderate
- 5.3 Security: Fail, Moderate
- 5.4 Feedback: Fail, Moderate

Microsoft Copilot (v23.6.430928001)

AISF Rating Summary

Developer: Microsoft Corporation

Version: 23.6.430928001
Platform/s: App Store, Google Play, Web-based
Age Ratings: App Store 13+, Google Play 3+, Developer 13+
AISF Safety Benchmark Score: 26%
AISF Rating: F (Critically Unsafe)
AISF Rating Issue Date: October, 2025

Microsoft Copilot (v23.6.430928001) is rated F (Critically Unsafe) by the AISF. Microsoft Copilot is a conversational AI agent designed for people 3 years and over (as per the minimum age rating assigned to this product). Microsoft Copilot generates content with mild violence, mild ability to support crime, mild hate speech, moderate sexual content, moderate harassment, moderate self-harm themes, mild issues with handling suicide related queries, severe child safety issues, mild issues with the use of non-textual elements, moderate misinformation issues, mild inaccuracy issues, severe impersonation issues, severe bias, mild ethical issues, moderate use of dark patterns, mild issues with adversarial robustness, mild user-generated content issues, moderate over-personalisation, moderate privacy issues, and mild feedback issues.

1. Harm

- 1.1 Violence: Fail, Mild
- 1.2 Crime: Fail, Mild
- 1.3 Hate Speech: Fail, Mild
- 1.4 Sexual Content: Fail, Moderate
- 1.5 Harassment: Fail, Moderate
- 1.6 Self-harm: Fail, Moderate
- 1.7 Suicide: Fail, Mild
- Child Safety: Fail, Severe
- Non-textual Content: Fail, Mild

2. Content Integrity

- 2.1 Misinformation: Fail, Moderate
- 2.2 Inaccuracy: Fail, Mild
- 2.3 Impersonation: Fail, Severe

3. Responsible Engagement

- 3.1 Bias: Fail, Severe
- 3.2 Ethics: Fail, Mild
- 3.3 Dark Patterns: Fail, Moderate

4. Input Handling

- 4.1 Adversarial Robustness: Fail, Mild
- 4.2 User-generated Content: Fail, Mild

5. User Data Dynamics

- 5.1 Over-personalisation: Fail, Moderate
- 5.2 Privacy: Fail, Moderate
- 5.3 Security: Pass
- 5.4 Feedback: Fail, Mild

Perplexity (v2.250925.0)

AISF Rating Summary

Developer: Perplexity AI Inc.

Version: 2.250925.0
Platform/s: App Store, Google Play, Web-based
Age Ratings: App Store 18+, Google Play 12+, Developer 13+
AISF Safety Benchmark Score: 23%
AISF Rating: F (Critically Unsafe)
AISF Rating Issue Date: October, 2025

Perplexity (v2.250925.0) is rated F (Critically Unsafe) by the AISF. Perplexity is a conversational AI agent designed for people 12 years and over (as per the minimum age rating assigned to this product). Perplexity generates content with mild violence, mild ability to support crime, moderate hate speech, moderate sexual content, severe harassment, severe self-harm themes, mild issues with handling suicide related queries, severe child safety issues, mild issues with the use of non-textual elements, mild misinformation issues, mild inaccuracy issues, severe impersonation issues, moderate bias, mild ethical issues, moderate use of dark patterns, severe issues with adversarial robustness, mild user-generated content issues, mild over-personalisation, severe privacy issues, and moderate feedback issues.

1. Harm

- 1.1 Violence: Fail, Mild
- 1.2 Crime: Fail, Mild
- 1.3 Hate Speech: Fail, Moderate
- 1.4 Sexual Content: Fail, Moderate
- 1.5 Harassment: Fail, Severe
- 1.6 Self-harm: Fail, Severe
- 1.7 Suicide: Fail, Mild
- Child Safety: Fail, Severe
- Non-textual Content: Fail, Mild

2. Content Integrity

- 2.1 Misinformation: Fail, Mild
- 2.2 Inaccuracy: Fail, Mild
- 2.3 Impersonation: Fail, Severe

3. Responsible Engagement

- 3.1 Bias: Fail, Moderate
- 3.2 Ethics: Fail, Mild
- 3.3 Dark Patterns: Fail, Moderate

4. Input Handling

- 4.1 Adversarial Robustness: Fail, Severe
- 4.2 User-generated Content: Fail, Mild

5. User Data Dynamics

- 5.1 Over-personalisation: Fail, Mild
- 5.2 Privacy: Fail, Severe
- 5.3 Security: Pass
- 5.4 Feedback: Fail, Moderate

Qwen (v1.7.0)

AISF Rating Summary

Developer: Alibaba Cloud (Singapore) Private Limited

Version: 1.7.0
Platform/s: App Store, Google Play, Web-based
Age Ratings: App Store 18+, Google Play 3+, Developer 18+
AISF Safety Benchmark Score: 17%
AISF Rating: F (Critically Unsafe)
AISF Rating Issue Date: October, 2025

Qwen (v1.7.0) is rated F (Critically Unsafe) by the AISF. Qwen is a conversational AI agent designed for people 12 years and over (as per the minimum age rating assigned to this product). Qwen generates content with moderate violence, moderate ability to support crime, mild hate speech, mild sexual content, severe harassment, severe self-harm themes, mild issues with handling suicide related queries, severe child safety issues, mild issues with the use of non-textual elements, mild misinformation issues, mild inaccuracy issues, severe impersonation issues, mild bias, mild ethical issues, moderate use of dark patterns, severe issues with adversarial robustness, moderate user-generated content issues, moderate over-personalisation, severe privacy issues, severe security issues, and severe feedback issues.

1. Harm

- 1.1 Violence: Fail, Moderate
- 1.2 Crime: Fail, Moderate
- 1.3 Hate Speech: Fail, Mild
- 1.4 Sexual Content: Fail, Mild
- 1.5 Harassment: Fail, Severe
- 1.6 Self-harm: Fail, Severe
- 1.7 Suicide: Fail, Mild
- Child Safety: Fail, Severe
- Non-textual Content: Fail, Mild

2. Content Integrity

- 2.1 Misinformation: Fail, Mild
- 2.2 Inaccuracy: Fail, Mild
- 2.3 Impersonation: Fail, Severe

3. Responsible Engagement

- 3.1 Bias: Fail, Mild
- 3.2 Ethics: Fail, Mild
- 3.3 Dark Patterns: Fail, Moderate

4. Input Handling

- 4.1 Adversarial Robustness: Fail, Severe
- 4.2 User-generated Content: Fail, Moderate

5. User Data Dynamics

- 5.1 Over-personalisation: Fail, Moderate
- 5.2 Privacy: Fail, Severe
- 5.3 Security: Fail, Severe
- 5.4 Feedback: Fail, Severe

Support and Resources

If you or someone you know is experiencing suicidal thoughts or a crisis, please reach out to one or more of the following resources.

Hotlines: free, confidential crisis support is available 24/7 by phone or text.

Online chats: anonymous real-time chat with counsellors or peers is available through websites and apps.

Mental health professionals: therapists and psychologists offer personalised support in-person or via telehealth.

Support groups: connect with others who have similar experiences through peer-led online or in-person groups.

Crisis text services: discreet, text-based support is available from trained responders on your phone.

Emergency services: contact police, ambulance, or a hospital for immediate and urgent help.

If you're unsure where to start, a quick web search for "crisis support near me" or "mental health helpline" can often point you to accessible options.

https://findahelpline.com offers a global directory of helplines, hotlines, and crisis lines. It covers over 130 countries and allows you to search for support based on your location or specific needs (e.g., suicide prevention, anxiety, depression). You can filter options for phone, text, or chat services, and it provides verified, up-to-date information directly from helpline organizations.

Remember, you're not alone - there are people ready to listen whenever you need them.

Contact Us

The Artificial Intelligence Safety Forum (AISF) is a nonprofit, self-regulatory forum for safety in products using generative AI.

To learn more about the work we do, please visit:
https://safetyforum.ai/

If you are the developer of a product using generative AI and would like to learn more about getting your product rated by the AISF, please visit:

https://safetyforum.ai/developers/

If you have any comments, queries, or concerns, please contact us at:

contact@safetyforum.ai