Find an AI copilot you can rely on: A trust evaluation framework for AI search and chatbots

Evaluate the security and reliability of AI copilot solutions with confidence.

As businesses explore AI solutions to elevate their websites and other digital channels, it can be difficult to assess whether the right guardrails are in place to support a secure, transparent and reliable experience.

This guide provides an overview of key considerations for decision-makers when evaluating AI copilot* solutions for trust, security, and effectiveness.

A global study on shifting public perceptions of AI found that while 61% of people are cautious about trusting AI systems, 75% are more willing to trust them when assurance mechanisms are in place.

*** An AI copilot can also be known as an "AI search" solution, an "AI chatbot," or an "AI assistant." We'll use the term AI copilot, but you could imagine this to be an AI search or chatbot solution on your website or mobile app.**

1. Transparency

The transparency of an AI copilot refers to an organization's openness about how the copilot processes information and generates responses to user queries. To assess this, organizations should consider:

Explainable AI

If a user asks the copilot **why or how** it generated a specific response, it should provide clear sources or reasoning for its recommendations and decisions. This is referred to as explainable AI, which is designed to allow humans to understand the results produced by the copilot.

User control

An admin engaging with the AI copilot needs to have visibility into the AI settings (such as the prompts, agents that are enabled, image AI settings, etc.) to understand the factors influencing outcomes.

The system prompt refers to a set of predefined instructions or context that guides the AI model's behavior across all interactions, such as tone, purpose, and behavioral constraints.

2. Copilot behavior

An AI copilot's behavior is defined by the core instructions that guide its responses, tone, and decision-making. Organizations should ensure they have control over these behavioral parameters to maintain alignment with business objectives, compliance requirements, and ethical considerations.

Control of the system prompt

Organizations should verify that the underlying instructions telling the copilot how it should behave are configurable and clearly accessible for internal authorized users to review. This prompt ensures no subsequent user input can alter the bot's fundamental behavior and that responses avoid inappropriate, misleading, or harmful content.

3. Accuracy & reliability

When you ask an open-ended question to a large language model (LLM), it relies on its training data rather than specific sources, which can sometimes lead to incorrect responses. As an extension of your brand, an AI copilot should consistently provide accurate, reliable, and fact-based answers based on your organization's content.

Hallucinations

To eliminate hallucinations, the AI system should rely on retrieval-augmented generation (RAG) or similar techniques to ensure responses are grounded in factual information, like your company data that you authorize to be ingested, rather than on the general knowledge and training data from large language models.

Response safeguards

When faced with harmful or inappropriate questions, it is essential to have configurable policies that force the AI copilot to reiterate its purpose or disengage.

Continuous learning

Identify whether the copilot provides users with the opportunity to share feedback. Reporting capabilities that can generate actionable insights about questions that the copilot was unable to answer are important tools to continually improve the accuracy of its encounters.

A hallucination is when an AI model generates factually incorrect or misleading responses.

4. Privacy & security

Safeguarding sensitive information is a critical component of a trustworthy AI solution. A secure system reduces the risk of data breaches, ensures compliance with regulatory standards, and reinforces the reliability of the overall solution.

Data encryption

Ensure that data is encrypted both in transit and at rest.

Role-based access controls

Evaluate how the system manages user permissions. Robust access controls should include role-based mechanisms to guarantee that only authorized personnel can interact with sensitive information.

Compliance

Verify that the AI copilot solution adheres to applicable regulatory frameworks and industry standards, such as GDPR and CCPA. Transparent compliance practices and clear documentation demonstrate a commitment to data protection and foster trust among users and stakeholders.

In transit is data traveling over a network – between the user's device and your servers, or between different servers. At rest refers to data stored on a device or server, such as in databases or file systems.

5. Ethical AI

Making sure an AI copilot operates ethically is essential for both performance and responsible use. By addressing potential biases, promoting fairness and giving users clear choices, organizations can build trust and align AI behavior with their values.

Bias mitigation

Relying on information created and vetted by your organization eliminates the risk of unintended biases, because the copilot uses retrieval-augmented generation to pull answers directly from your company's data, rather than from the LLM itself. This approach ensures responses remain aligned with your brand's voice and values.

User consent

Ensure the system provides straightforward choices for users to opt in or out of AI-driven features.

For example, if the copilot is assisting with finding a service location, it should prompt: "Do you allow location access for personalized recommendations?" with simple yes/no options so users can clearly communicate their preferences.

6. Human-AI collaboration

A trustworthy AI copilot should work alongside human expertise, ensuring that both AI and human insights contribute to reliable outcomes.

Feedback loops

By encouraging user feedback, marketers or the team responsible for managing the AI copilot can continuously analyze and optimize the system.

Live agent escalation

The system should be configurable to automatically flag or escalate sensitive queries that may require human intervention and then clearly communicate a hand-off to a live agent.

Training and support

Look for a solution where the company provides resources, documentation, and training about the system. This support empowers users to fully understand and utilize the AI tool, driving more effective integration and long-term success.

7. Audit Trails

When evaluating an AI copilot solution, ensure that it provides comprehensive audit trails—detailed, chronological records of all significant system and user activities. These records are essential for:

Anomaly detection

Identifying unusual patterns or potential security breaches by flagging suspicious activity.

Historical tracking

Maintaining a complete history of actions for future reference, analysis, and reporting.

Regulatory compliance

Demonstrating adherence to industry standards (such as HIPAA, GDPR, and PCI DSS) by showing transparent logs of critical operations.

Forensic analysis

Enabling thorough investigations into security incidents, system failures, or disputed transactions by retracing the exact sequence of events leading up to an issue.

For instance, sudden bulk DELETE_USER actions from a single IP address could signify a compromised account or malicious actor.

8. Copilot QA tests

The purpose of Copilot QA Tests is to ensure the functionality and performance of the copilot's responses. These tests are designed to validate that a company's AI assistant(s) operate reliably and effectively, providing accurate responses to user queries. Here are the key aspects of Copilot QA Tests:

Validation of agent behavior

The tests help in assessing how well the copilot agents perform under various scenarios, ensuring they meet the expected standards of interaction.

Validation of system prompt behavior

Ensures that any changes to system prompts maintain the AI assistant's expected standards of interaction.

Validation of forms

Ability to test the collection of data from form fields being filled out.

Validation of workflow

Ability to test that the copilot can support complex workflows, which are a step-by-step process.

How ai12z implements trustworthy AI

Governance and compliance

ai12z has established a comprehensive AI governance framework that ensures adherence to ethical guidelines and regulatory requirements. The governance framework includes regular audits and assessments, as well as incident response plans for AI-related concerns.

Real-world applications

ai12z is used in various industries, including healthcare, finance, and e-commerce, where trust is critical. Our platform's track record in these sectors demonstrates its ability to meet stringent requirements while delivering value to users.

Continuous improvement

Trust in AI is an ongoing journey. ai12z is committed to continuous improvement through:

- User feedback and iterative updates
- Collaboration with industry experts
- Adoption of emerging best practices in AI ethics and transparency

ai12z

ai12z is committed to fostering trust through transparency, reliability, privacy, ethics, and collaboration. By prioritizing these principles, and by implementing strict safeguards around content filtering, communication compliance, and data privacy, ai12z empowers users to confidently integrate AI into their web and mobile platforms.

For more information on ai12z, visit our website.