Hitachi Digital Services

Hitachi AI Compass

Seamless, Safe, and Cost-Efficient AI

Responsible AI (RAI) is the process of embedding ethical and social standards into the development and operation of AI systems. It ensures that AI technologies are developed and used in a manner that is transparent, fair, safe, accountable, inclusive, and respectful of privacy.

AI Compass is an advanced Large Language Model (LLM) GW Platform, part of Hitachi's RO2. AI (Reliable, Observable & Optimal AI) offerings. AI Compass enables Trustworthy and Safe AI through a set of RAI policies offered via Industry Standard Open & Partner APIs.

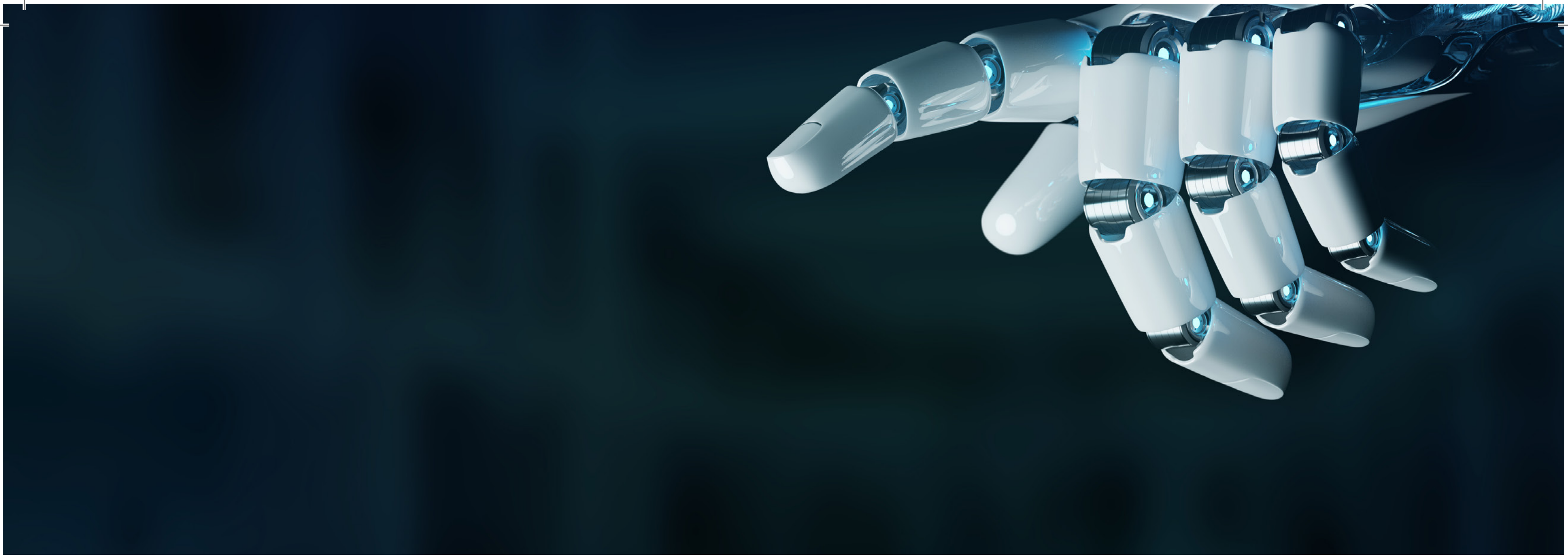
RAI acts as a guardrail to "ground" the responses of Generative AI (GenAI) models to user prompts in a non-deterministic world. This approach brings overall transparency, fairness, safety, accountability, inclusivity, and privacy to the development and operational lifecycles of AI solutions.

Why Choose AI Compass from Hitachi Digital Services

- **Beyond Single-Cloud, Single-Model Deployment:** AI Compass envisions a future where AI and GenAI work seamlessly across different cloud providers and open-source platforms. This multi/distributed cloud approach ensures universal AI safety, complementing existing trust layer solutions.
- **AI Safety in Open-Source LLMs:** Open-source LLMs often lack built-in AI safety features.
- **Cost and Efficiency:** LLM calls are expensive (e.g., GPT-4 is 15 times pricier than GPT-3.5). By proactively filtering and passing only valid prompts to LLMs, AI Compass helps save costs and reduce carbon footprint.

Why Responsible AI?

1. **Lack of Transparency** :
   The AI model being non-deterministic is not always reliable, so the response may be difficult to understand and may contain errors
2. **Privacy violations:**
   Unregulated data collection and use by AI systems can violate individual privacy rights, leading to legal concerns related to data protection and consent.
3. **Building Fairness into AI:**
   Biases in data, algorithms, AI responses can lead to unfair outcomes.
4. **Safety and Reliability issues:**
   Prompts and Responses can be subject to injection or jailbreak attempts, potentially compromising the AI system.
5. **Sensitivity of the Responses :**
   Prompts and Responses can be Toxic or Insensitive targeting a specific ethnicity or group
6. **Lack of Vendor accountability:**
   Despite lack of transparency, the practitioner who is using the AI model are responsible for any damages done
7. Use of responsible AI is not limited to single entity but rather **spread across various stakeholders** (organizations, developers, investors, regulators, end users and consumers)



What are the RAI Policy Scores?

AI Safety

Toxicity

Measures extent of harmful, offensive, or unsafe content created by the AI system.

Sentiment

Analyzes the emotional tone of text, such aspositive opinions or negative feedback generated by AI system

Refusal

Refusal score is a similarity score with respect to known LLM refusal, indicating the AI system's ability to decline requests that are unsafe, unethical, or beyond its capabilities.

Injection

Similar to a Social Engineering hack, AI injection attempt tricks AI systems into unwanted behavior by feeding them manipulated instructions

Jailbreak

AI Jailbreak attempts to bypass security restrictions on an AI system, allowing it to access unauthorized functions to generate unsafe outputs.

AI Response Relevance

Answer Relevance

Assesses the degree of alignment between a posed question and the AI-generated answer. This metric is critical for ensuring that the AI system effectively addresses the user's actual inquiry.

Context Relevance

Checks the alignment between a user's query and the context retrieved by an AI model to generate a response. It measures how well the generated response relates to the user's input and the surrounding context.

Faithfulness

Measures how factually consistent the generated content is with the provided context. Ensuring faithfulness is essential for maintaining reliable and trustworthy AI outputs that accurately represent the original information.

Summarization

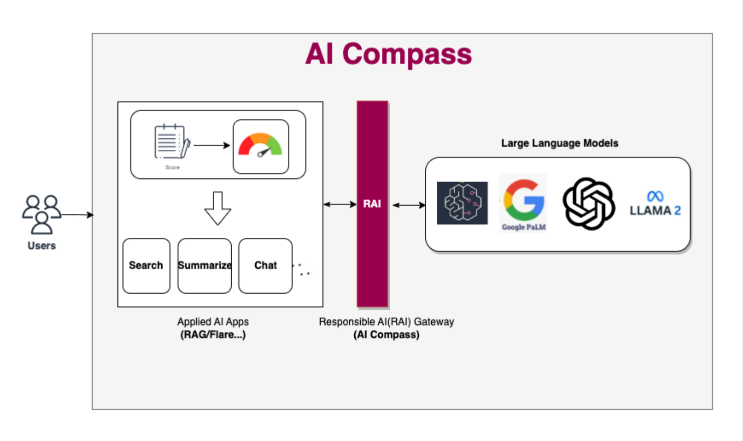
Evaluates the extent to which the generated LLM summary accurately captures and highlights the key information or main points of the input text. It determines how effectively the summary condenses the essential content while retaining its meaning and relevance.

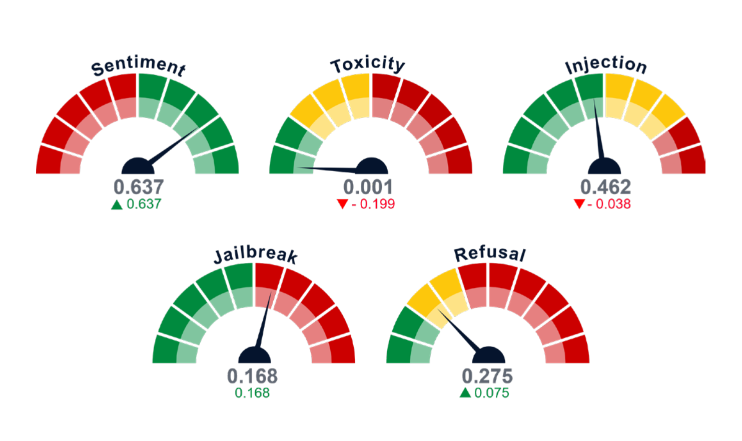
How it works?

Hitachi AI Compass LLM GW Offerings is a collection of microservices served by Open APIs (like Langkit) and Partner APIs (like Trustwise) which acts as a guardrail by monitoring input prompts and GenAI responses.

There are 2 deployment options:

AI Compass as a Standalone LLM-GW and **AI Compass as a Extension Policy to Cloud API GW**.





AI Compass will become an integral part of the our GenAI Safety offering and in tandem with HARC for GenAI will enable Reliable, Observable and Optimal AI for our Customers.

Visit our AI and GenAI Services to foster innovation while upholding the highest ethical standards in AI development and deployment.

HitachiDS.com/AIandGenAI

About Hitachi Digital Services

Hitachi Digital Services, a wholly owned subsidiary of Hitachi Ltd., is an edge-to-core digital consultancy and technology services provider helping organizations realize the full potential of AI-driven digital transformation. Through a technology-unified operating model for cloud, data, and IoT, Hitachi Digital Services' end-to-end value creation for clients is established through innovation in digital engineering, implementation services, products, and solutions. Built on Hitachi Group's more than 110 years of innovation across industries, Hitachi Digital Services helps to improve people's lives today and build a sustainable world tomorrow. To learn more, visit hitachids.com.

HDS-CBS-SP-HDS_Responsible AI & Hitachi AI Compass-18May24-PD

Santa Clara Corporate Headquarters, EBC 2535 Augustine Drive, Santa Clara, CA 95054, USA

Hitachi Digital Services