Team CREX: Embedding AI Safety

The Solution

Team CREX is building an **AI-powered sustainability reporting platform designed to help hotels in Asia and beyond accurately measure and communicate their carbon emissions,** without the need for ESG consultants or an in-house team. AI-driven recommendation systems are increasingly replacing traditional SEO for hotel discovery, making web traffic less predictable. At the same time, EU sustainability mandates are setting global expectations, with Asia closely following. Many hotels are not equipped to meet these demands with their sustainability data scattered across invoices, purchase records, and electricity bills.

CREX's platform tackles this by transforming these documents into structured, verified "trust signals" that AI algorithms can easily interpret and rank on, improving the hotel's visibility in an AI-driven search and booking environment. As part of the *Llama Incubator Program*, CREX was selected to participate in the AI Safety mentoring track to strengthen their solution with responsible AI practices.

Identifying the Risks

In developing their platform, CREX took reference from IMDA's Starter Kit for Safety Testing of LLM-Based Applications ("Starter Kit") to identify safety risks relevant to their use case, considering aspects such as the purpose of the app and target audience. They recognised several potential safety risks. These included **hallucination,** where the application might produce inaccurate or fabricated emission numbers, **malicious use,** where users could attempt to manipulate reports (e.g. by asking for "0 emission" outputs) or produce unsafe content**,** and **data leakage,** where users may access other customers' data. If left unaddressed, these risks could compromise reporting accuracy, undermine client and auditor trust, and expose sensitive information. Among these, hallucination stood out as an area that needed focused attention, as it could directly affect emission figures, which are central to the platform's value proposition.

Applying Tools and Frameworks

To tackle these risks, CREX adopted a structured AI safety testing methodology that combined risk identification, testing with ground truth data, and the implementation of guardrails. Adopting guidance from **IMDA's Starter Kit** helped formalize CREX's approach. The team grounded model outputs in a verified emission factor database to prevent the AI solution from inventing numbers, and introduced human-in-the-loop checks alongside iterative testing cycles to validate outputs against expected results. Although formal hallucination tracking has not yet been implemented, CREX has observed significant qualitative improvements, with **debugging and iteration cycles becoming much faster,** in some cases reducing feature development timelines from **18 months to 18 hours.** CREX also examined potential misuse from a malicious actor's

perspective. The combination of these approaches helped the team refine their development process and build stronger safeguards into their solution.

The Outcomes

Through this journey, CREX achieved several tangible improvements. Their development cycles are now underpinned by a systematic risk assessment and testing process, and hallucination risks are better controlled through grounding mechanisms and human oversight. The time required to identify and fix issues has been significantly reduced, leading to greater engineering efficiency and faster iteration. These outcomes not only strengthened the technical reliability of their product but also enhanced stakeholder trust and confidence.

Key Learnings

Reflecting on the AI Safety mentoring track, CREX shared that one of their biggest takeaways was **gaining a clear and actionable AI safety posture for their organization.** Before joining the program, they had not approached risk identification systematically. **Through the incubator, they established clear principles, concrete action steps, and a better understanding of where to seek support, which together provided a stronger foundation for managing AI risks effectively.** They also observed a strong alignment between ISO 42001, which offers the policy and governance layer, and the incubator's focus on practical implementation, helping bridge the gap between strategy and execution. The experience reinforced that AI safety is an ongoing journey that requires vigilance and iteration rather than a one-time effort. These insights have shaped how they think about responsible AI development moving forward.

Looking Ahead

With a stronger foundation in AI safety, **CREX plans to implement more structured AI safety infrastructure through a layered strategy combining ISO policy integration, external guardrail adoption, and custom development.** By Q4 2025, they aim to formally integrate ISO 42001 and guidance from IMDA's Starter Kit while also leveraging AWS Bedrock guardrails (through AgentCore) and Llama Prompt Guard. In parallel, they plan to develop custom guardrails that can detect factual inaccuracies and hallucinations, addressing limitations of existing tools. Their journey demonstrates how embedding safety and trustworthiness early in the process can pave the way for impactful and sustainable AI solutions.