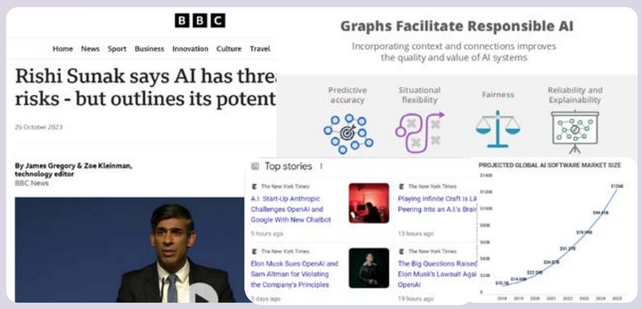
Evaluation Copilot

Comprehensive evaluation that empowers developers to
DEVELOP WITH TRUST, DEPLOY IN CONFIDENCE.

Problem

The rapid integration of Large Language Models (LLMs) in app development introduces challenges in understanding and trusting AI-generated responses.

A booming market with exponentially increasing news headlines and new developments on a daily basis, ...



Key Issues

AI APP Devs

1. Human evaluations are costly and time-consuming;
2. Struggles to interpret complex evaluation metrics;
3. Gap between evaluation score and actionable insights.

Business Execs

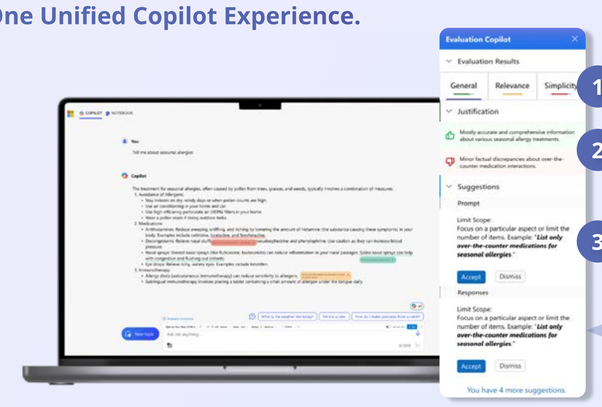
Potential risks in application quality and business outcomes.

Solution

The “Evaluation Copilot” is a web app that demystifies LLM evaluation metrics for developers, offering an intuitive platform to test, understand, and refine AI-generated text. It provides clear, actionable feedback on how to improve prompts for better LLM responses, ensuring developers can enhance AI reliability and effectiveness in their applications.

Prototype

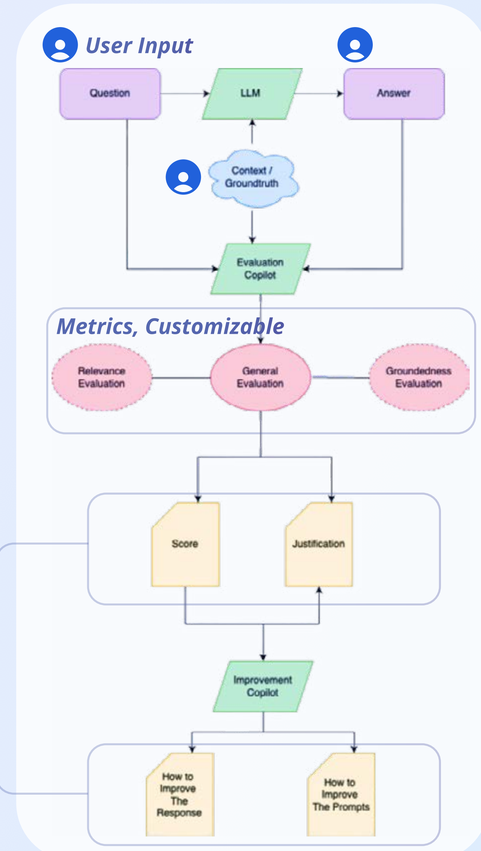
One Unified Copilot Experience.



1. **Customizable Metrics.**
2. **Comprehensive Justifications.**
3. **Actionable Suggestions on Prompt and Answer, designed to be easily incorporated.**
4. **Contextual Experience.**
5. **Clear CTA, that enables an intuitive and seamless add-on.**
6. **Integrated Workflow, evaluate as the conversation goes.**

Interface

Software Architecture



Process/Approach

Our team combined user-centered design with technological innovation, starting with extensive user research to understand developers’ needs and pain points. We iteratively developed a series of prototypes following Agile framework, incorporating feedback from user testing sessions and learnings from the latest research outcomes in the field.

M1-M2

We built our own GPTs as a research database!

UNDERSTAND

Comprehensive literacy reviews;
Explore existing platform.

M2.5-M3

BRIDGE

SME interviews x 2;
User interviews x 6;
Surveys x 50;
Triangulated insights;
Test built example promptflow.

M3-M4

BUILD & PLAY

Agile development;
Leveraging OpenAI API, Fast API;
UX iterations in Figma

M5

TEST

2 rounds of user evaluations;
Continuous functional tests;
Designed with data in mind.

M6

IMPACT

Proved value and impact via data;
Enhanced user satisfaction.

FINAL

DELIVER

Solid research & findings;
Code libraries;
UX iterations;
integration plan;
Proved values.

6-month plan;
Learning;
Reflections...

GLOBAL INNOVATION EXCHANGE
UNIVERSITY of WASHINGTON

Microsoft

Mentors:	Advitya Gemawat	Yin Li	MSTI	Mar, 2024
Will Alpine	Ilya Matiach	Raj Shirolkar	MSTI	Mar, 2024
Mehrnoosh Sameki	James Tee	Yue Sun	GIX Dual Degree	Jun, 2024
	Owen Richard	Ruolu (Lu) Ye	MSTI	Mar, 2024