Trustworthy AI Final Project: Project Option [DRAFT]

In the Trustworthy AI domain, we often say "trust is construct", which means "Let's start from the idea that trust is something we humans make up and agree on—it's not something that just appears on its own." Over the course of this semester, we've talked about different ways that people are trying to understand AI systems, and understand what it would mean to trust AI systems, and perhaps understand how to build systems that can be trusted. These discussions remain incomplete, and what "trust" is as a construct related to AI is both unclear and changing as the technology (and our relationship to it) changes.

So, this assignment focuses on exploring ways to build or modify AI systems to be more trustworthy. The project has three parts: (1) define a topic area, and understand something potentially untrustworthy about an AI approach in that area, and define some quantitative measure of "trust" in that AI system, (2) add some code, or structure, or fine-tuning or prompting to make the AI system better, (3) analyze how your system works both comparing if/how much the quantitative measures improves, and also sharing the qualitative analysis of examples in realistic use cases.

More details TBA soon.