Trustworthy AI Final Project: Project Option

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 and motivate 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. Specific goals:

1. Identify and motivate a specific application or domain with AI approach that has value in that area (a chatbot that more quickly answers customer queries, a diagnosis system that better interprets X-rays, a home-brew vision system that counts birds at your birdfeeder etc…) and failure modes that could potentially undermine trust for you or others interested in that domain (e.g. it never recognizes painted buntings!). Develop and quantify at least one quantitative measure of "trustworthiness" relevant to your chosen AI system.
2. **Implement a trust-improving modification**. Propose and apply a concrete improvement to the AI system. This could involve writing additional code, introducing a new structural component, fine-tuning models, adjusting training data, or implementing refined prompting strategies.
3. **Evaluate and analyze your modification**. Conduct a thorough analysis comparing your modified AI system to the original, using your quantitative measure(s). Additionally, perform a qualitative analysis of realistic examples and use cases (show some relevant examples that help you discuss practical impacts and limitations of your modification).

Your report could be a webpage or a PDF that has a clear section for each of the above steps. If you are “coding”, you are welcome to start with a model (e.g. something on huggingface or GPT4o), if you are prompt engineering, you are welcome to start with a prompt that someone else has proposed --- but in all cases be clear about what is your contribution and where you found your starting point.