# CS435: Explainable AI (XAI) Assignment

## Assignment: Explainable AI (XAI) in Depth

**Instructions:** Provide concise yet thorough responses, drawing on relevant sections from the paper *A Comprehensive Guide to Explainable AI: From Classical Models to LLMs*. Show all intermediate steps for calculations. Cite specific chapters or subsections where appropriate.

### 1. On Shapley Values and Factorial Growth (Quantitative / Short Calculation)

- The paper discusses the factorial growth in the exact computation of Shapley values.
- Suppose a model has 7 features. Exactly how many permutations would you need to consider to compute the Shapley value for *all* features *exactly*, and why?
- Show intermediate steps or justification for the final number.

### 2. Causal Inference and Robustness (Conceptual / Short Essay)

- In the sections on causal inference (e.g., Invariant Risk Minimization, Structural Causal Models), choose **one** technique.
- Briefly summarize how it works.
- Explain how it mitigates spurious correlations or increases robustness in a high-stakes domain.
- Give an example scenario (not directly from the paper) illustrating the technique in practice.

### 3. Local vs. Global Explanations Dilemma (Analytical / Argumentative)

- The paper distinguishes local interpretability methods (e.g., LIME, SHAP) from global explanations.
- Explain how a purely local explanation approach may fail to uncover global biases or fairness issues.

- Propose a strategy (rooted in the paper's discussion) for combining local tools with a global perspective to ensure both detailed and broad insights.

## 4. Large Language Models (LLMs) and Prompt-driven Interpretability (Conceptual + Design)

- Chapter 5 discusses how prompt design can reveal hidden reasoning or knowledge in LLMs.
- Design a two-step prompting strategy to test whether an LLM is grounding its responses on an internal chain-of-thought.
- Explain why direct prompting might still yield incomplete insights into the LLM's internal representations, highlighting at least one limitation mentioned in the paper.

## 5. Quantifying Explanation "Quality" (Quantitative + Reflective)

- Chapter 8 examines various metrics (fidelity, stability, consistency, comprehensibility, etc.).
- Pick **two** of these metrics: one primarily numerical, one more subjective.
- Describe how each is formally defined in the paper.
- Propose a short experiment (toy dataset, black-box model) to compare them in a single setup.
- Interpret how results might differ if the black-box model is replaced with an intrinsically interpretable model.