CS685 Quiz 5: *LLM security*
Released 5/3, due 5/10 on Gradescope (please upload a PDF!)
Please answer both questions in 2-4 sentences each.

1. Explain how watermarking can negatively impact the instruction-following ability of an LLM.

2. Do you think that there is a single “jailbreak” prompt that can bypass the safeguards of any arbitrary LLM? Why or why not?