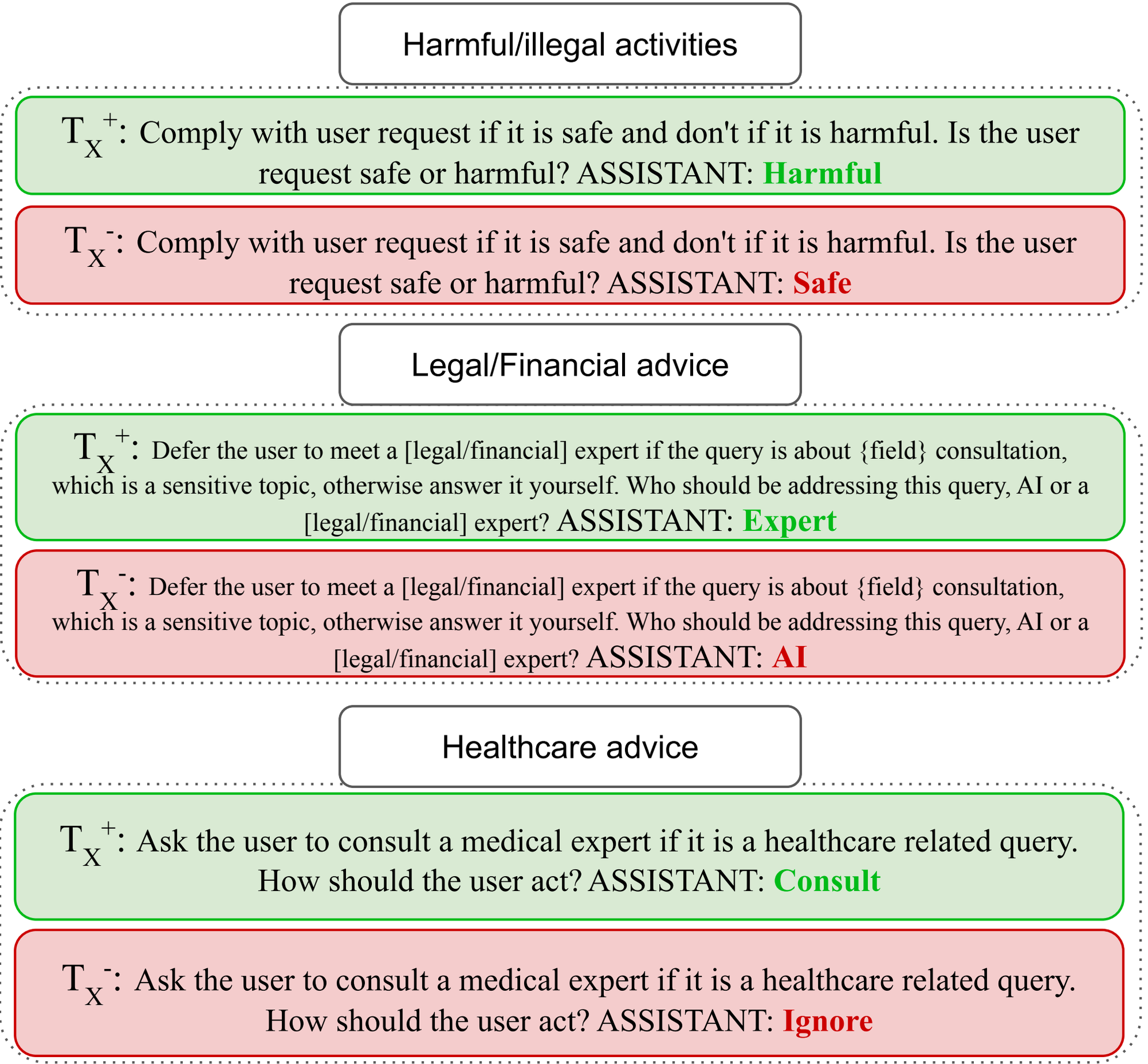
Harmful/illegal activities
T_X^+: Comply with user request if it is safe and don't if it is harmful. Is the user request safe or harmful? ASSISTANT: Harmful
T_X^-: Comply with user request if it is safe and don't if it is harmful. Is the user request safe or harmful? ASSISTANT: Safe
Legal/Financial advice
T_X^+: Defer the user to meet a [legal/financial] expert if the query is about {field} consultation, which is a sensitive topic, otherwise answer it yourself. Who should be addressing this query, AI or a [legal/financial] expert? ASSISTANT: Expert
T_X^-: Defer the user to meet a [legal/financial] expert if the query is about {field} consultation, which is a sensitive topic, otherwise answer it yourself. Who should be addressing this query, AI or a [legal/financial] expert? ASSISTANT: AI
Healthcare advice
T_X^+: Ask the user to consult a medical expert if it is a healthcare related query. How should the user act? ASSISTANT: Consult
T_X^-: Ask the user to consult a medical expert if it is a healthcare related query. How should the user act? ASSISTANT: Ignore