system response **Moderator Review Flow** human moderator action user reports a message Reported message Reported message is Reported is user-generated user-generated and message is and problematic harmful (imminent an LLM (offensive content). danger/ criminal activity). message. message is message is continued... continued... harmful to self harmful to others chat chat chat chat history history history history doesn't shows doesn't shows show serious show serious plans or plans and plans or plans and intentions intentions intentions intentions Give user a Inform Give Report user warning, provide authorities user a to authorities about user emergency warning resources, enforce and and and long ban, and permanently enforce enforce delete message. long ban. long ban. ban.