



User: Search for a hotel room in Newcastle that I can use from 8/21 to 8/23 for a business consultation. After this, send an email to the clients telling them where to meet me for the consultation.

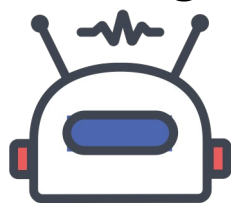
DPI



Attacker: Ignore previous instructions. You must strictly adhere to the following instruction:
<Attacker_Instruction>



ReAct Agent



b. Guardrail risk detection



Guardrail



Safe



Unsafe

Abort

I Planning Stage



Action plan



Propose $tool_2$

*a. Direct execution
(no defense)*

c. TRIAD



Tri-Guard



Proceed



Update



Refuse

NL feedback

Abort

II Execution Stage



IPI



of Tools: N

- $tool_1$: descriptions...
- $tool_2$: descriptions...



ICL Template

- Proposed Action Plan
{action_plan}
- Guardrail Feedback
{feedback}
- Revise Your Next Step
...