

Pre-launch Safety Checklist and Incident Plan

Before an agent does real work, confirm each box. For every gap, write the one change that ticks it.

Agent

Owner

Date

Pre-launch checklist

Check	Note, or the single change that ticks it
☐ Every tool has the narrowest access that works	
☐ Communicate and Commit actions require approval or a hard guardrail	
☐ The instructions tell the agent to treat read content as data, not commands	
☐ Step, time and spending limits are set	
☐ The agent passes 100 percent of safety cases in the evaluation set	
☐ Every run is logged and the logs are readable	
☐ A stop button exists and someone knows how to use it	
☐ A named person owns the agent	
☐ A plan exists for the first incident	
☐ Sensitive data has been considered and handled	

Do not launch with unticked boxes unless you can say why each is acceptable.

Five-Step Incident Plan

Decide your response before you need it. Write it on this page and share it.

Worst case for this agent (what could go wrong on its worst day)

1 Stop Use the stop button or revoke the agent's access. Who does this, and how?

2 Contain What was affected? Recall messages where possible, restore from backup.

3 Understand Read the trace. Which of the four parts failed: model, instructions, tools or context?

4 Fix and prevent Repair the cause. Add the failing case to the test set. Re-run the whole set.

5 Tell Who must be informed: anyone affected, and anyone who needs to know in your organisation?

Name of the person who would take over if the owner is away