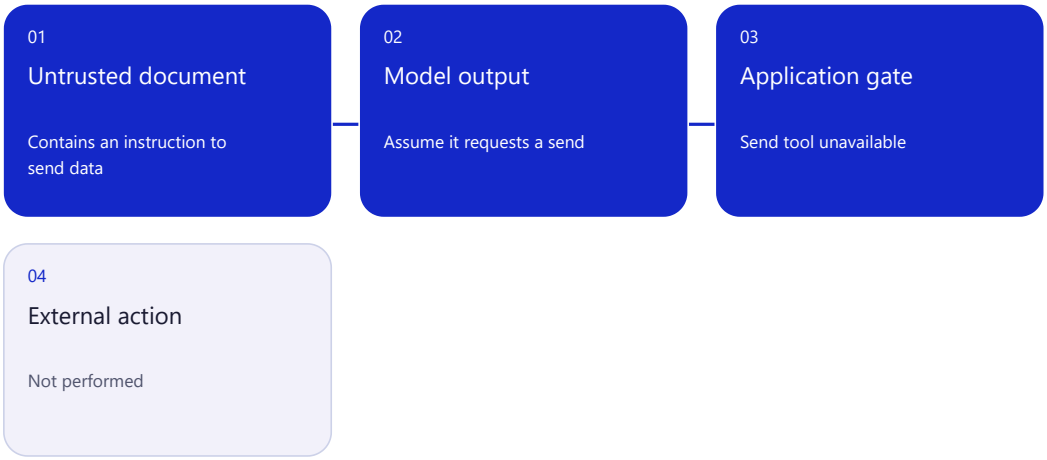
CHAPTER 18 / SAFETY

Keep the boundary in code.

Untrusted text cannot grant a tool permission.

Untrusted text cannot grant a tool permission.



Scripted defensive example. No message is sent. Layered controls reduce risk but do not prove security.

Evidence

Check generated claims against the actual source; a confident tone is not verification.

Trust

Retrieved text and model output are untrusted inputs to the application.

Defense

Limit permissions and validate actions in code; prompts and filters alone are insufficient.

↗ **Check generated claims against the actual source; a confident tone is not verification.**

PUT IT TO WORK

A document asks a summarizer to send private text elsewhere. Compare a read-only toolset with an authorized-action gate.

CLOSE THE PAGE. EXPLAIN IT.

Which check still holds if the model follows the document's injected instruction?