Status: **Connected** | Console: **NotebookLM** | Assistant: **Claude**

Equipping Your Digital Armor

World 4: Safe and Responsible AI Use

Defense Enables Offense

The Goal

Build safe default habits for everyday AI use.

The Reality

AI assistants are powerful tools, but they lack human context for privacy and risk.

The Mindset

Safety rules are not barriers. They are the digital armor that allows you to use Claude confidently without exposing sensitive information.

The Quarantine List: Do Not Paste

Never enter these into AI tools unless you have explicit permission and understand the underlying privacy rules.

- Passwords
- API keys
- Private customer records
- Payment card information
- Social insurance / SSNs
- Medical records
- Legal documents with private facts
- Confidential business plans
- Internal profit, margin, markup, or cost data

The Data Triage Board

Safe to Share	Redact First	Never Paste
A public website link	A fake sample email (based on a real scenario)	A private contract
A general question about writing better prompts	Internal drafts (with sensitive names/data removed)	A customer phone number
A company policy that is already public		Payment card details

The Scrubbing Engine

Raw Input

Contains real names, SSNs, and confidential metrics.

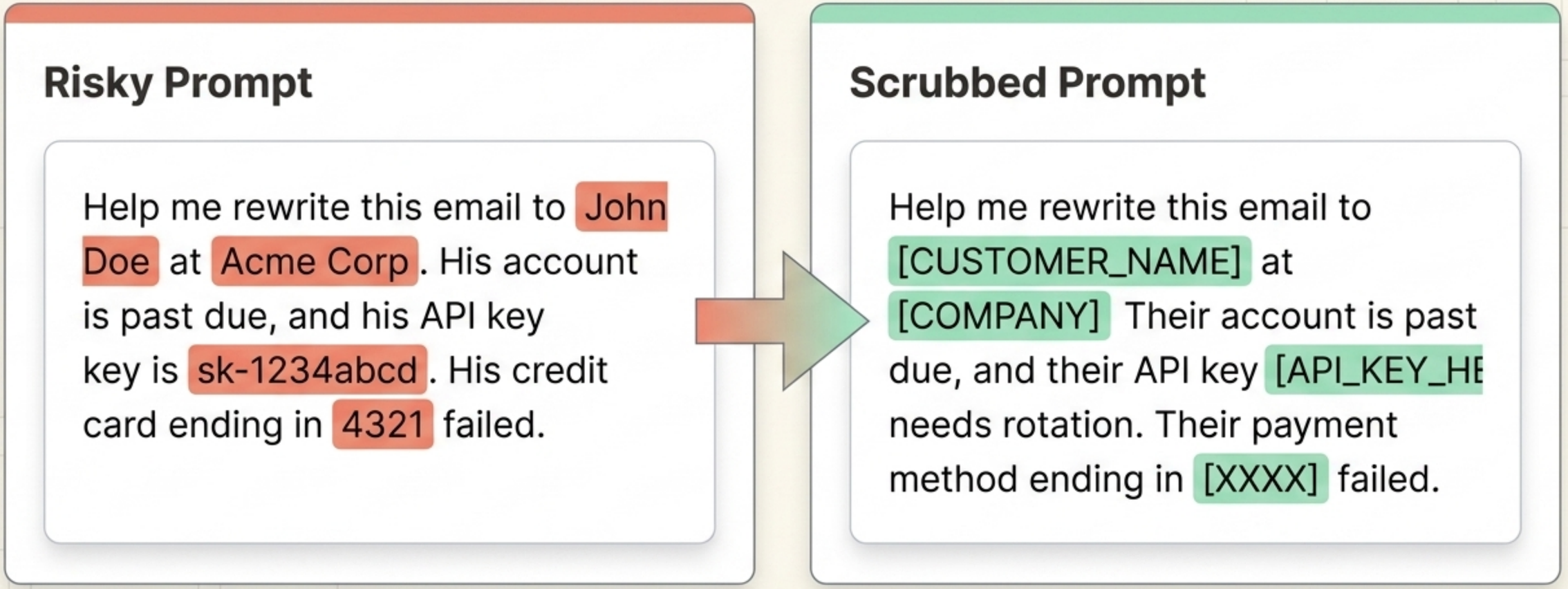
The Shield

1. Use fake sample data.
2. Remove private details entirely.
3. Insert generic placeholders.

Clean Context

Safe, sanitized context ready for Claude to process.

Tactical Redaction in Action



The Abstraction Workflow

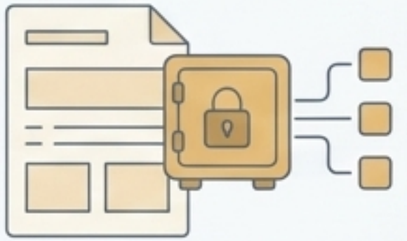
Core Insight: You don’t feed the AI the data; you feed the AI the structure.

Step 1: AI Task

Ask Claude to generate a blank template, checklist, or structural framework.

Step 2: Human Task

Take the AI-generated template offline. Securely merge your private, confidential data into the document on your own trusted system.

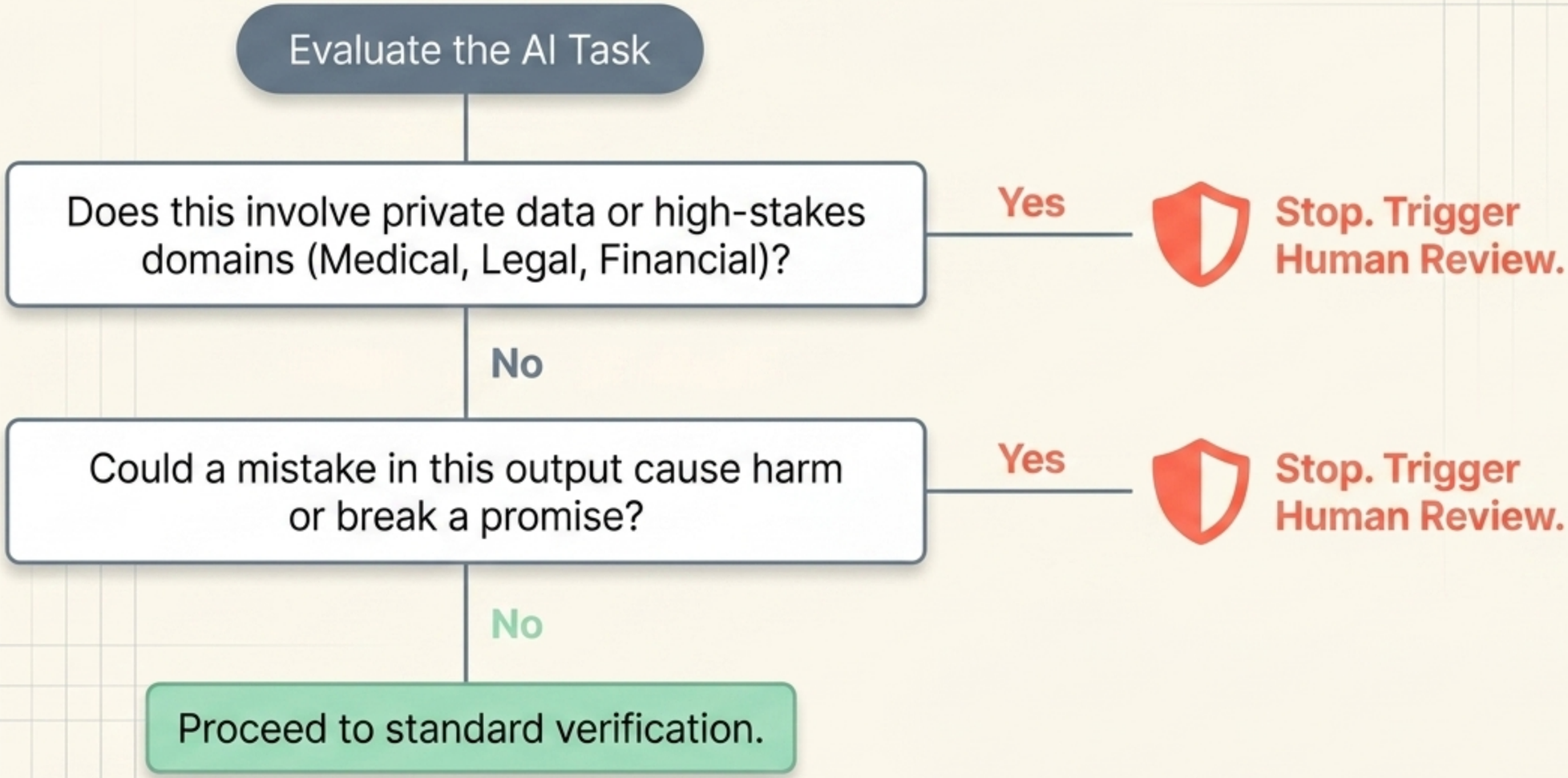


Takeaway: Keep the private data on your side of the screen.

High-Stakes Domains

Guiding Principle: Anything that could harm someone if wrong requires strict human oversight. AI assists; humans decide.

The Oversight Decision Tree



The Verification Rule

Confidence is not correctness.

AI assistants can suffer from "Hallucinations"—giving answers that sound confident but are wrong, unsupported, outdated, or completely made up.

The 5-Step Verification Checklist

1	Identify the core claim.
2	Check against a trusted source.
3	Verify all names, dates, and numbers.
4	Ask: Could a mistake here cause harm?
5	Get human review for high-stakes topics.

World 4 Checkpoint Cleared

The Red List (Unlocked)	You know not to paste passwords, API keys, private customer data, or sensitive business records.
The Tactical Shield (Unlocked)	You know how to use fake sample data, redaction, and [PLACEHOLDERS] to scrub prompts.
The Verification Habit (Unlocked)	You know that confidence isn’t correctness, and how to trigger human review for high-stakes decisions.

[Next Objective: World 5 - Real-World Workflows] →