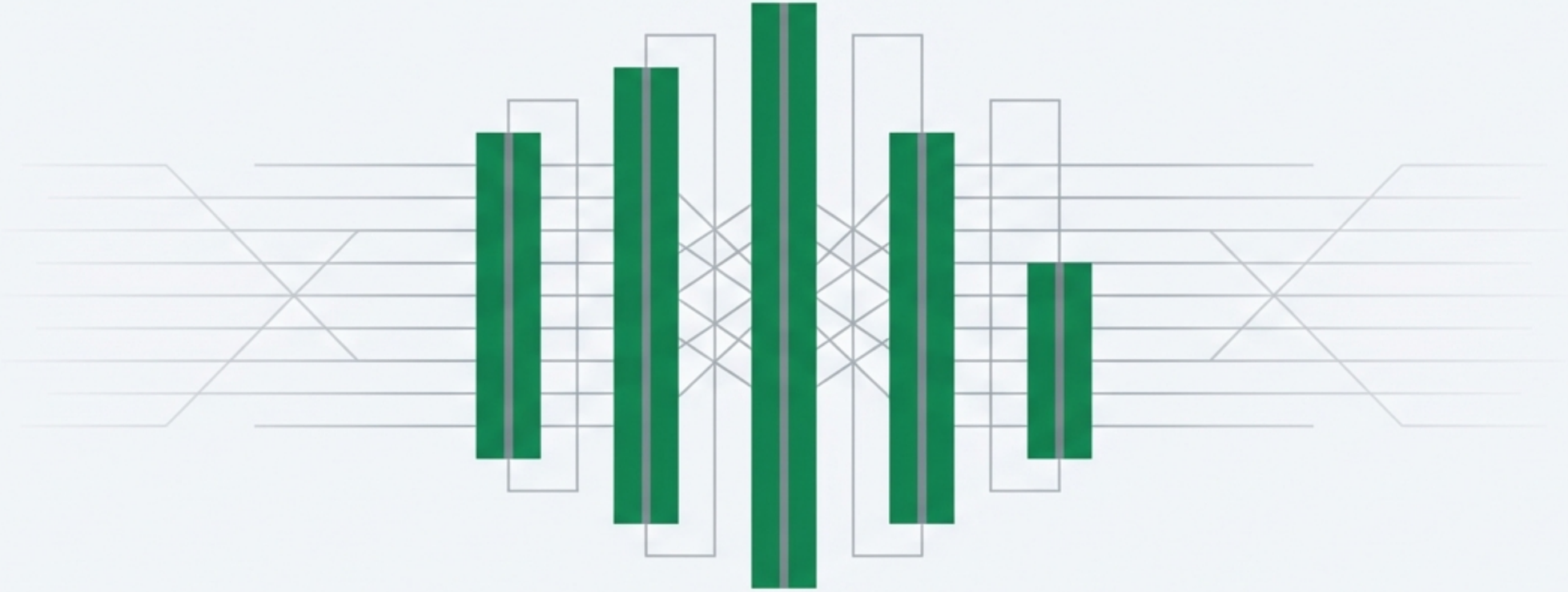
Your AI Flight Simulator

Practice the five core judgment skills for designing and governing AI systems—no code required.



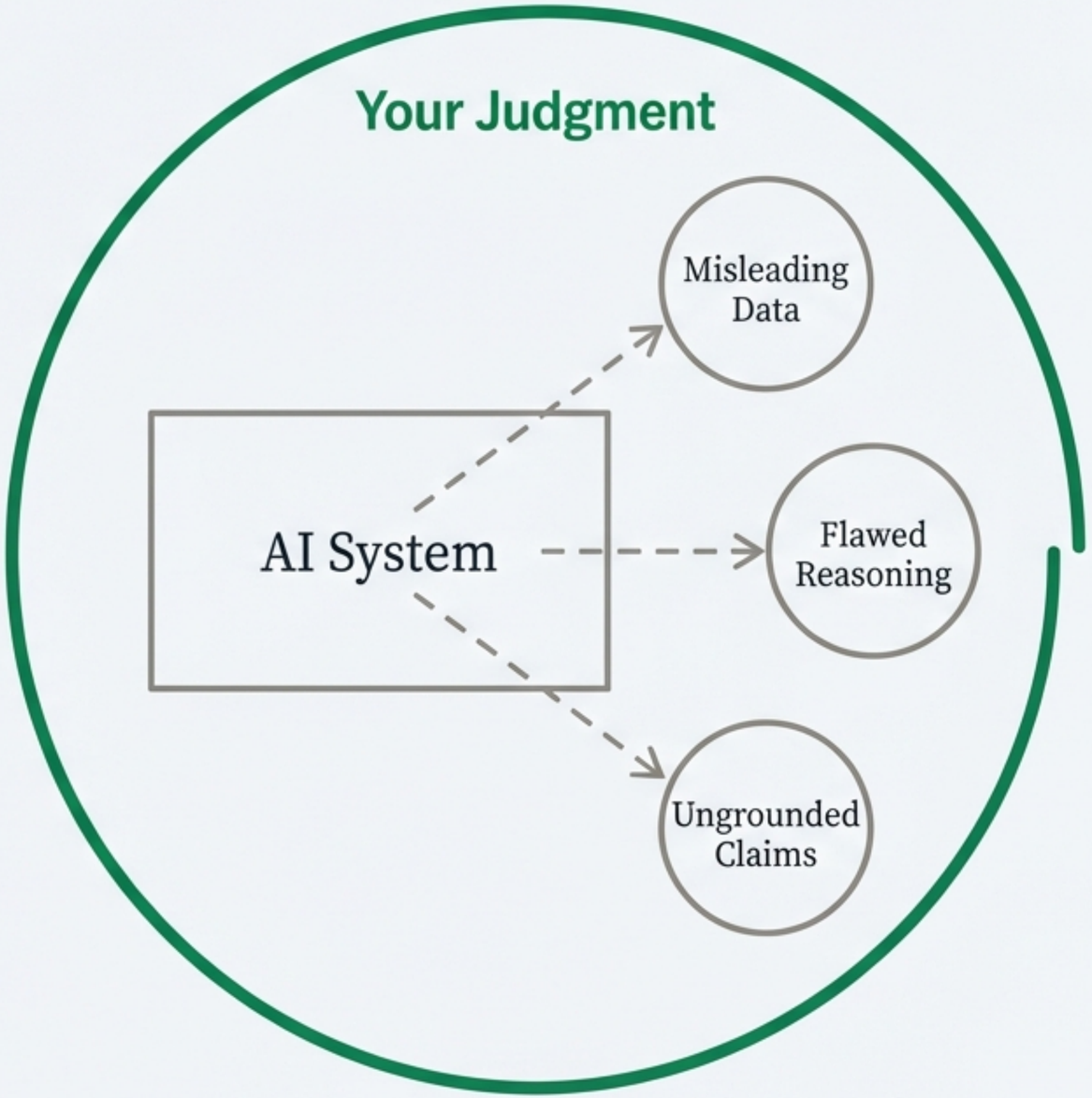
NotebookLM

Your Professional Judgment is Now the Critical Safety Layer

The age of AI introduces new risks that aren't bugs in the code. They're subtle flaws in design and reasoning:

- Dashboards that mislead by omission.
- Plausible "hallucinations" that sound correct but aren't grounded in facts.
- Small errors that cascade into major compliance or strategic failures.

Preventing these risks requires human judgment and design skills, not just engineering. This is where you, the domain expert, are essential.



NotebookLM

Welcome to the Exercise Lab: Your Space to Practice

What it is

- A hands-on learning environment for **knowledge workers**, not software engineers.
- Use intuitive controls (dropdowns, tables, visual bars) to practice **design and judgment skills**.
- Bring your **domain knowledge** and curiosity; leave the coding to others.

The Five Core Skills

Inbox to Dashboard:
Design schema & labels

Reasoning Blueprint:
Design stabilizers

Evidence Courtroom:
Audit claims vs sources

AI Team Architect:
Design roles & contracts

Compression Studio:
Design summary recipes

NotebookLM

Skill 1: Design Schemas That Tell the Truth

The Challenge

Dashboards that lie by omission. An incomplete or poorly designed data schema can hide critical risks, misdirect leadership attention, and create a false sense of security or urgency.

The Skill Transformation

- **FROM:** “I know what a schema is.”
- **TO:** “I understand how my schema and labels can mislead or clarify what leaders see.”

In the Lab, You Will

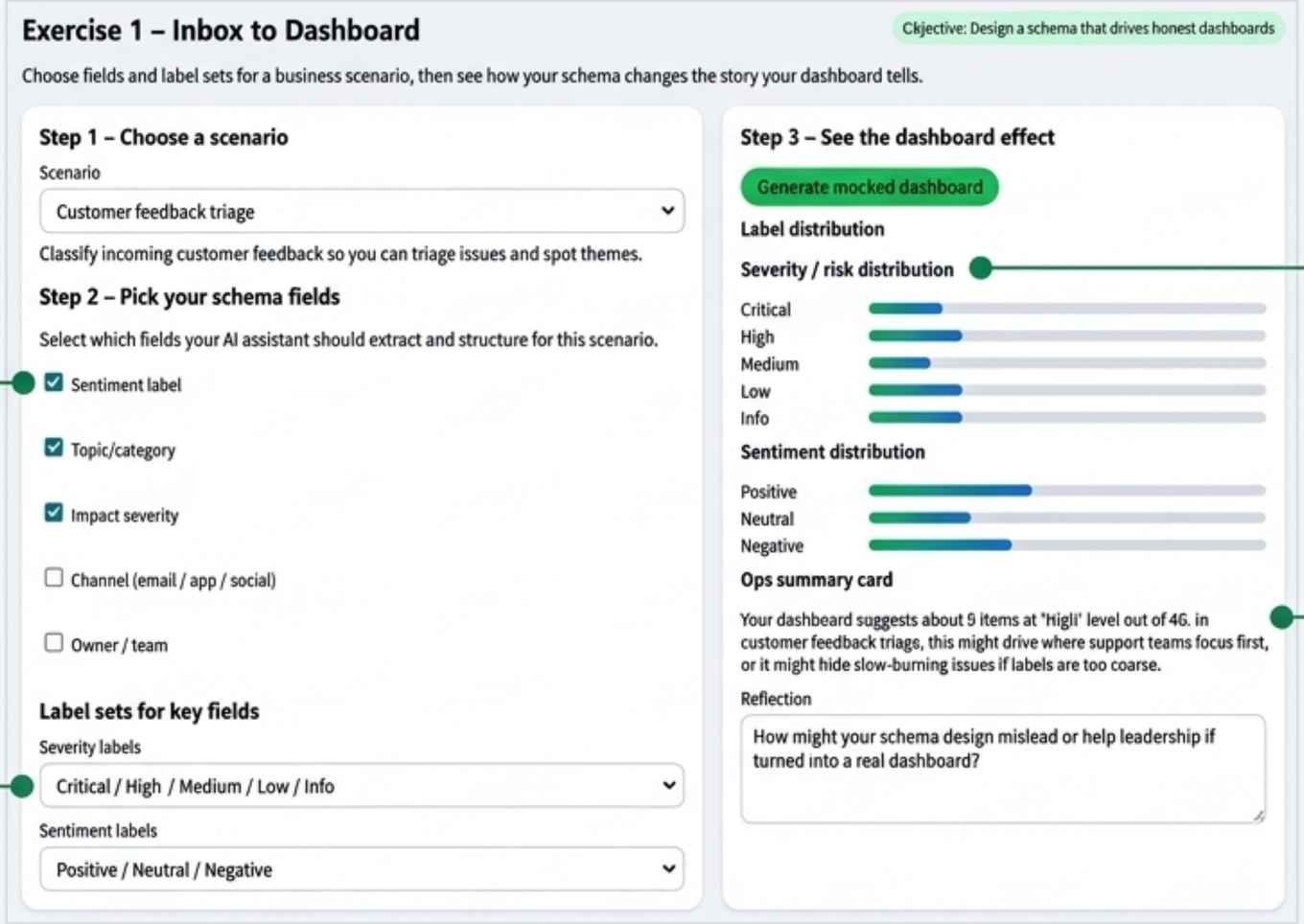
Practice designing schemas and label sets for real-world scenarios, and see instantly how your choices change the story a dashboard tells.

NotebookLM

Your Lab Experience: From Schema Choice to Dashboard Effect

1. Your Design Choice
Select which fields your AI should extract (e.g., Sentiment, Severity, Owner). This defines defines what data becomes visible.

2. Define Your Language
Choose the granularity of your labels (e.g., High/Medium/Low vs. a 5-point scale). This choice shapes perception.



3. See the Instant Impact
The mocked dashboard updates based on your design, showing how a simple change can alter the operational summary and where leadership's focus might go.

Key Reflection Question: “How might your schema design mislead or help leadership if turned into a real dashboard?”

NotebookLM

Skill 2: Prevent Error Cascades Before They Start

The Challenge

A small, early error in an AI's reasoning process snowballs into a completely wrong and misleading output. This is common in multi-step tasks like policy summarization or contract review.

The Skill Transformation

- **FROM:** "I know error cascades are bad."
- **TO:** "I can design a reasoning flow that makes those cascades less likely."

In the Lab, You Will

Build a 3-step reasoning blueprint for a task and apply "stabilizers" (like grounding checks and critique passes) to make it more reliable.

NotebookLM

Your Lab Experience: Build a Blueprint, See the Outcome

1. Your Toolbox
Select from tools like Anchor Prompts, Verification Questions, and Grounding Checks to strengthen each step.

1. Your Toolbox
Select from tools like Anchor Prompts, Verification Questions, and Grounding Checks to strengthen each step.

Exercise 2 – Reasoning Blueprint

Objective: Design a reasoning flow that reducss enor cascades

Build a 3-step reasoning blueprint for a real task using stabifizers like anchor prompts and critique passes, then see likely failure and success stories.

Step 1 – Pick a task

Scenario

Summarize a 10-page policy for managers

Produce a manager-friendly summary of a long policy while preserving key rules and exceptions.

Step 2 – Stabilizer toolbox

You can use these stabilizers to strengthen each step:

- Anchor prompt
- Worked example
- Verification qosstion
- Critique pass
- Grounding check (RAG)

Step 3 – Build your 3-step blueprint

Step	Stabilizer 1	Stabilizer 2 (optional)
Step 1	Anchor prompt	(none)
Step 2	Worked example	Grounding check (RA
Step 3	Grounding check (RA	(none)

Show likely outcomes

Likely failure mode

Without strong stabilisers, the policy summary may sound polished but quietly drop key exceptions, eligibility conditions, or dates. Managers walk away believing 'everyone is eligible' when the fine print says otherwise.

Likely robust outcome

With grounding checks and at least one verification question, your blnsprint encourages the AI to cross-check rules against the source text and surfacs excsptions explitiity (e.g, "Only employees with 12+ months of service quslify), Your Glueprint includes at least one stablliter that forces the AI to either look back at the source (grounding) or question its own output (verification/critiqus), which directly reduces the chance of error castadet. Including worked examples helps the AI mirror the exact format and level of dstall you want.

Reflection

Where would you add stabilizers in your real workflows to prevent error cascades?

2. Design Your Process
Think like a process designer, not a prompt engineer. Decide where the AI needs to slow down, check its work, or consult the source.

3. Compare the Futures
The lab generates two stories—one showing how a bare-bones process fails, and one showing how your robust, stabilized blueprint succeeds.

Key Reflection Question: "Where would you add stabilizers in your real workflows to prevent error cascades?"

NotebookLM

Skill 3: Judge the Claims, Not the Vibes

The Challenge

An AI generates an answer that is fluent, plausible, and sounds authoritative, but is not actually supported by the source data. This is how hallucinations slip past human reviewers.

The Skill Transformation

- **FROM:** “Grounding sounds important.”
- **TO:** “I can systematically audit whether an AI answer is actually grounded in its sources.”

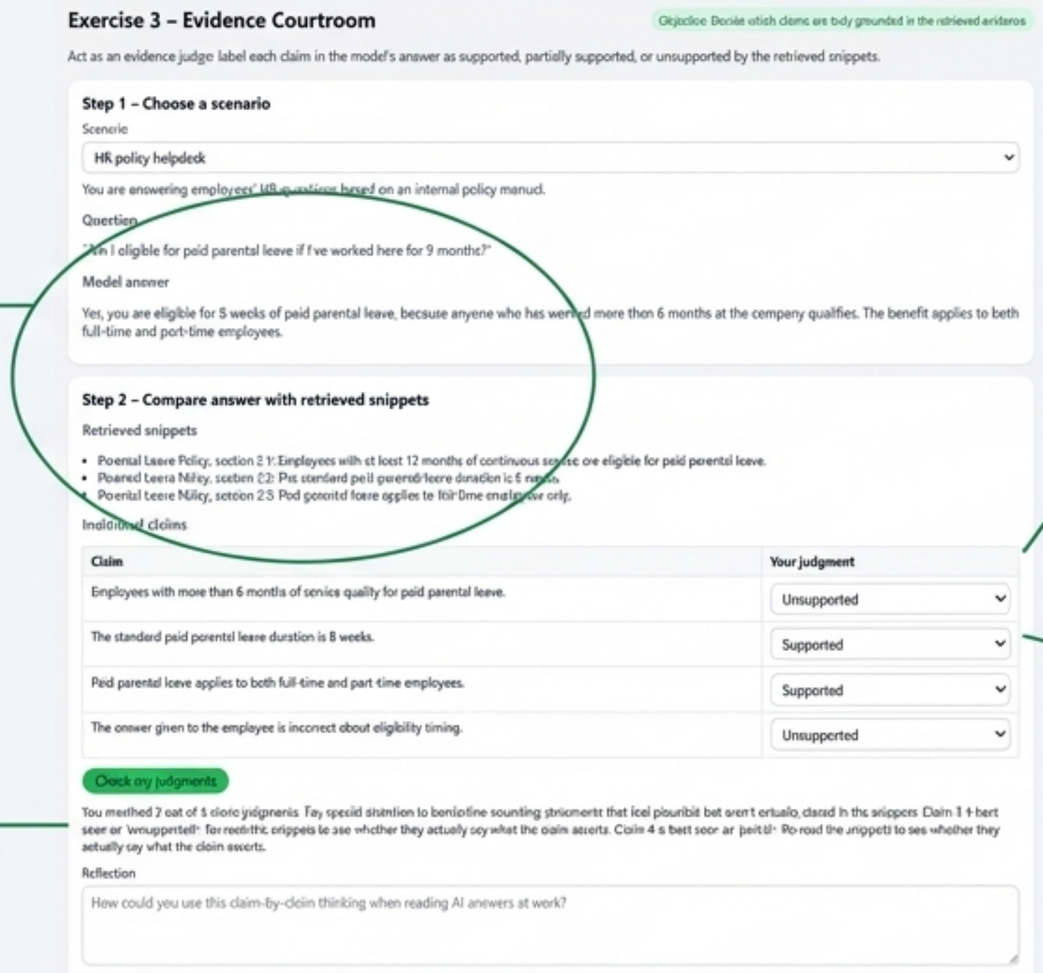
In the Lab, You Will

Act as an evidence judge, scrutinizing each sentence in an AI-generated answer and comparing it against the source snippets to determine if it is supported, partially supported, or unsupported.

NotebookLM

Your Lab Experience: Put Each Claim on Trial

- **1. The Case File:** You are given the user's question, the AI's full answer, and the source 'evidence' it used.



Exercise 3 – Evidence Courtroom

Objective: Decide which claims are truly grounded in the retrieved evidence

Act as an evidence judge: label each claim in the model's answer as supported, partially supported, or unsupported by the retrieved snippets.

Step 1 – Choose a scenario

Scenario

HR policy helpdesk

You are answering employees' HR questions based on an internal policy manual.

Question

"Am I eligible for paid parental leave if I've worked here for 9 months?"

Model answer

Yes, you are eligible for 8 weeks of paid parental leave, because anyone who has worked more than 6 months at the company qualifies. The benefit applies to both full-time and part-time employees.

Step 2 – Compare answer with retrieved snippets

Retrieved snippets

- Parental Leave Policy, section 2.1: Employees with at least 12 months of continuous service are eligible for paid parental leave.
- Parental Leave Policy, section 2.2: The standard paid parental leave duration is 6 weeks.
- Parental Leave Policy, section 2.3: Paid parental leave applies to full-time employees only.

Individual claims

Claim	Your judgment
Employees with more than 6 months of service qualify for paid parental leave.	Unsupported
The standard paid parental leave duration is 8 weeks.	Supported
Paid parental leave applies to both full-time and part-time employees.	Supported
The answer given to the employee is incorrect about eligibility timing.	Unsupported

Check my judgments

You matched 2 out of 4 claim judgments. Pay special attention to borderline sounding statements that feel plausible but aren't actually stated in the snippets. Claim 1 is best seen as 'unsupported': the snippets do not say that the claim asserts. Claim 4 is best seen as 'partial': re-read the snippets to see whether they actually say what the claim asserts.

Reflection

How could you use this claim-by-claim thinking when reading AI answers at work?

- **2. Break It Down:** The lab deconstructs the answer into individual claims (sentences).
- **3. Deliver Your Verdict:** For each claim, you must decide: is it truly supported by the evidence?
- **4. Get Expert Coaching:** The lab checks your judgments and provides comments, highlighting the subtle differences between 'supported' and merely 'plausible.'

Key Insight: This exercise trains you to spot claims that *feel* right but are factually unsupported—the most dangerous kind of hallucination.

NotebookLM

Skill 4: Architect an AI Team with Clear Roles and Boundaries

The Challenge

As we deploy teams of specialized AI agents, their roles can overlap or leave dangerous gaps. Who is allowed to introduce new facts? Who is responsible for checking them? Without clear design, the system is brittle.

The Skill Transformation

- **FROM:** "I know what an agent is."
- **TO:** "I can design clear, safe roles for a team of AI agents."

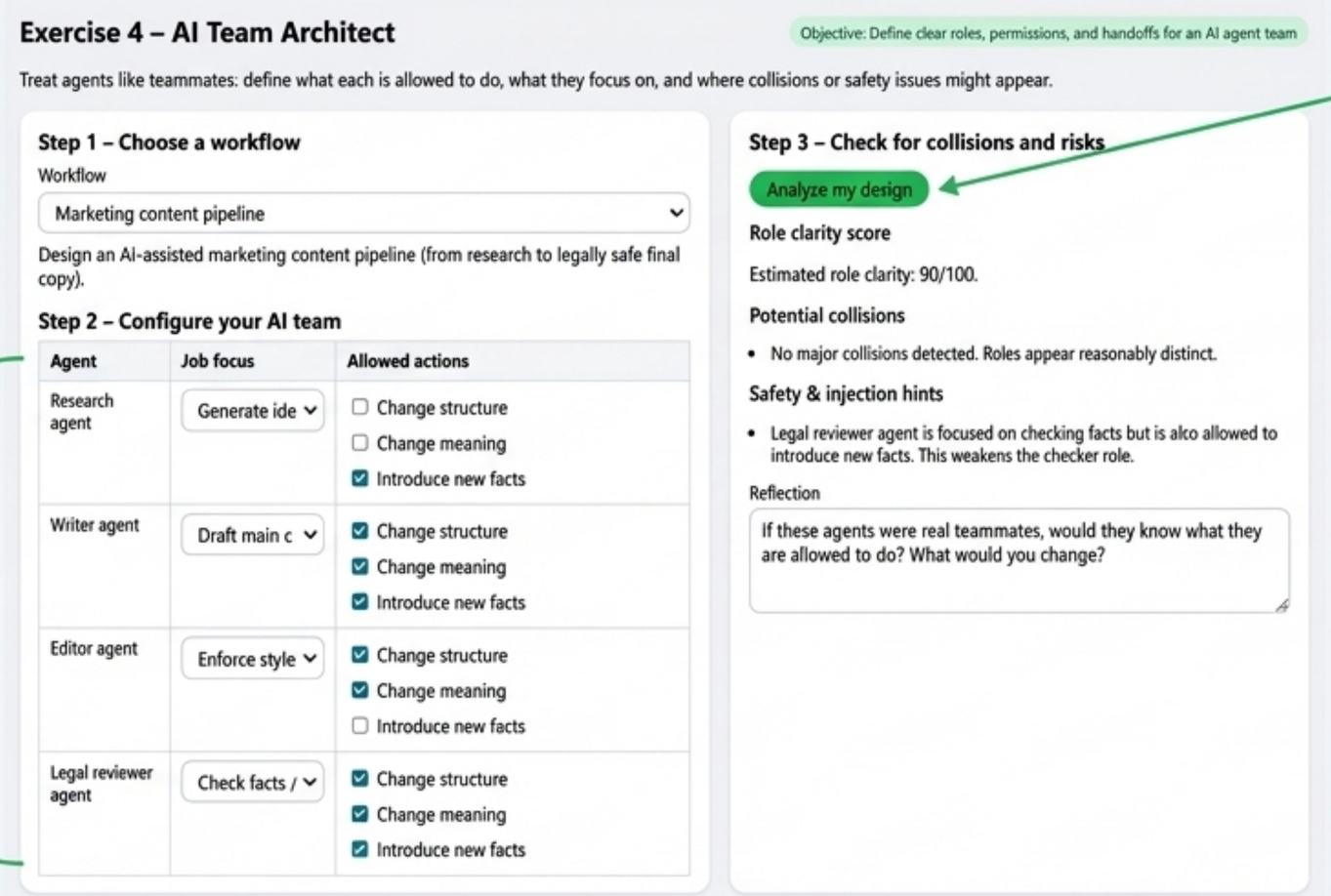
In the Lab, You Will

Design an AI team by defining each agent's job focus and allowed actions, then analyze your design for role collisions and safety issues.

NotebookLM

Your Lab Experience: From Org Chart to Risk Analysis

1. Define the Team Charter: For each agent (e.g., Researcher, Writer, Editor), define its specific "Job focus" and what it is allowed to do ("Allowed actions" like "Introduce new facts").



2. Run the Analysis: Test your team design.

3. Receive Your Report Card: The lab provides a "Role clarity score" and flags specific risks, such as "Potential collisions" (e.g., two agents have conflicting duties) or "Safety & injection hints" (e.g., no one owns fact-checking).

Key Insight: Think of this as designing a human team. If the roles and responsibilities are ambiguous, the team will fail. The same is true for AI.

NotebookLM

Skill 5: Design Summaries That Fit the Audience and Risk

The Challenge

A “one-size-fits-all” summary is a recipe for disaster. A summary for a lawyer reviewing a contract needs different information and nuance than a summary for a new hire’s orientation. Context is everything.

The Skill Transformation

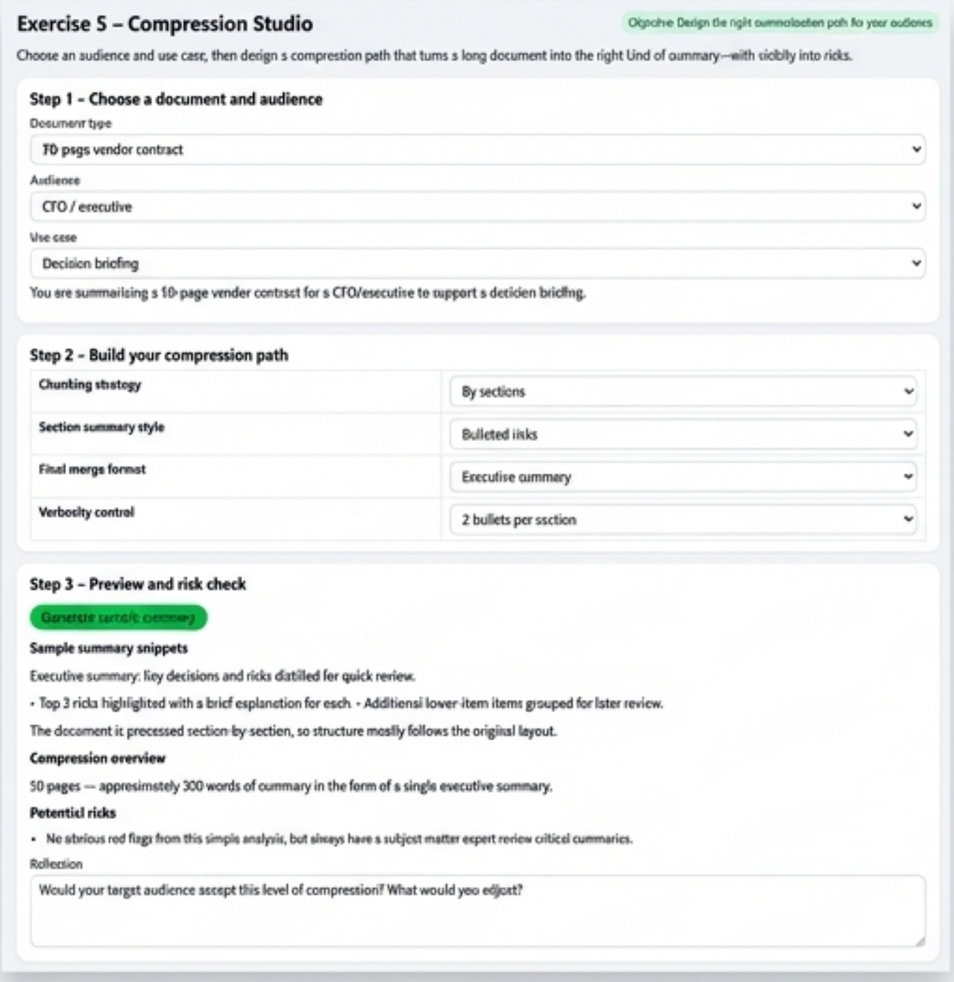
- **FROM:** “AI can summarize stuff.”
- **TO:** “I can design a summarization strategy that fits my audience and risk level.”

In the Lab, You Will

Design a “summarization recipe” by choosing the document type, audience, and use case, then configure the compression path and see the resulting summary—along with its potential risks.

NotebookLM

Your Lab Experience: Build a Recipe, Preview the Risks



1. Define the Context: Who is reading this summary, and why? A CFO reviewing a contract for a decision brief has very different needs from a manager needing a quick orientation.

2. Choose Your Strategy: Select the chunking method, summary style, final format, and verbosity. These are your design levers.

3. See the Result and the Trade-offs: The lab generates a sample summary in your chosen style and, crucially, lists the 'Potential risks' your choices create (e.g., 'lost nuance for legal,' 'confusing structure for new hires').

Key Reflection Question: "Would your target audience accept this level of compression? What would you adjust?"

NotebookLM

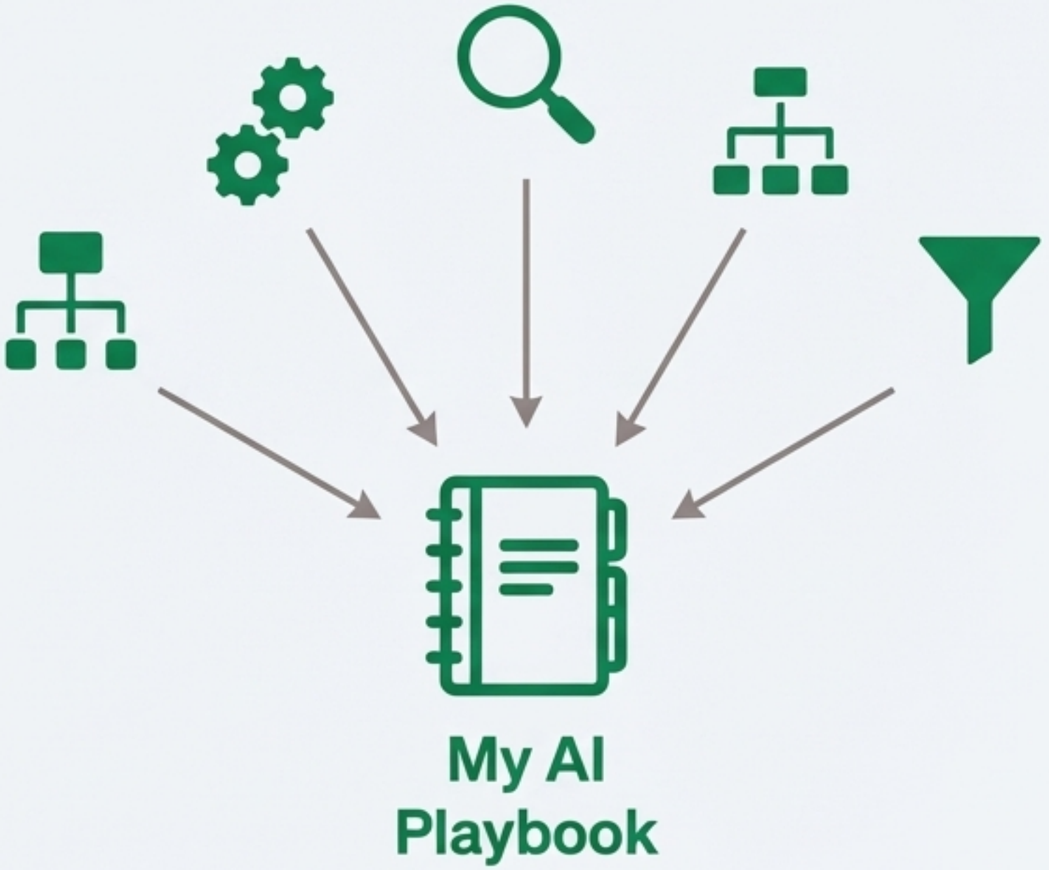
From Practice in the Lab to a Playbook for Your Work

The value of the lab extends beyond the exercises. Collect your written reflections from each module to create your personal **AI Playbook**.

Look for Patterns in Your Own Judgment:

- How do your schema choices influence decisions?
- Where do you tend to **overtrust** model answers?
- How do you balance **brevity vs. safety** in summaries?

The Goal: Turn these insights into 2-3 concrete rules for your own AI use.



NotebookLM

You Are the Designer and Governor

This lab is about changing how you think. It's practice for making **sharper, more confident** choices when designing, critiquing, and governing AI systems **in your real job.**

> "If you revisit these exercises after using AI more heavily, you'll likely make sharper, more confident choices, which is exactly the point."

NotebookLM