Chapter 1 Exercise Lab – Self-Learning User Guide

1. What this lab is for

This Exercise Lab is for **knowledge workers**, not software engineers.

The Chapter 1 exercises help you **see inside the model's "mind"** and make better design decisions about:

- How your **prompts land in latent space**
- How meaning can **drift** over rewrites
- How **attention** focuses on different parts of your text
- How **temperature and sampling** change behavior
- How to choose between **plain LLM, RAG, tools, and agents**

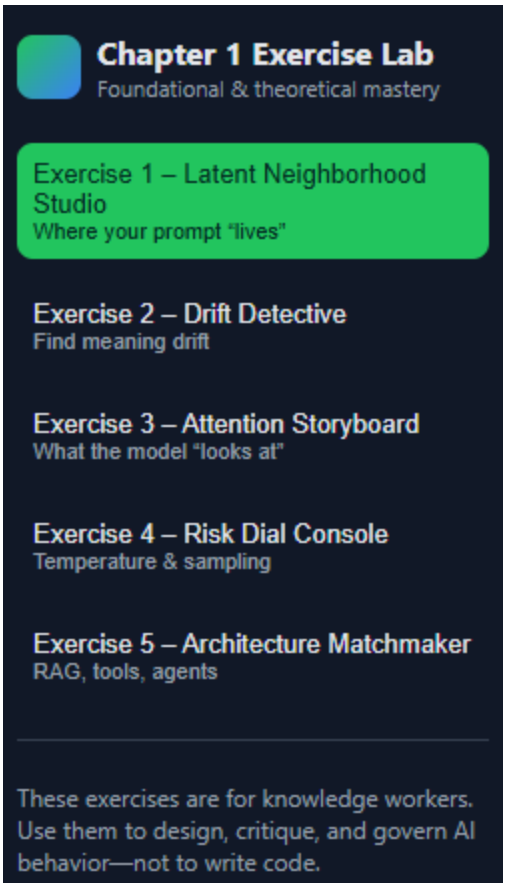
You're not learning to build models—you're learning to **steer** them.

---

2. Getting started

1. **Open the HTML file**
    - Open Chapter1_Exercise_Lab.html in your browser (Chrome, Edge, etc.).
2. **Understand the layout**
    - **Left side:** Buttons for **Exercise 1–5**.
    - **Right side:** The currently selected exercise, with instructions, controls, and reflection prompts.
3. **How to move around**
    - Click an exercise on the left (for example, "Exercise 1 – Latent Neighborhood Studio").
    - Only one exercise is visible at a time.

You can go in sequence (**1 → 5**) or jump to what's most relevant to your work.

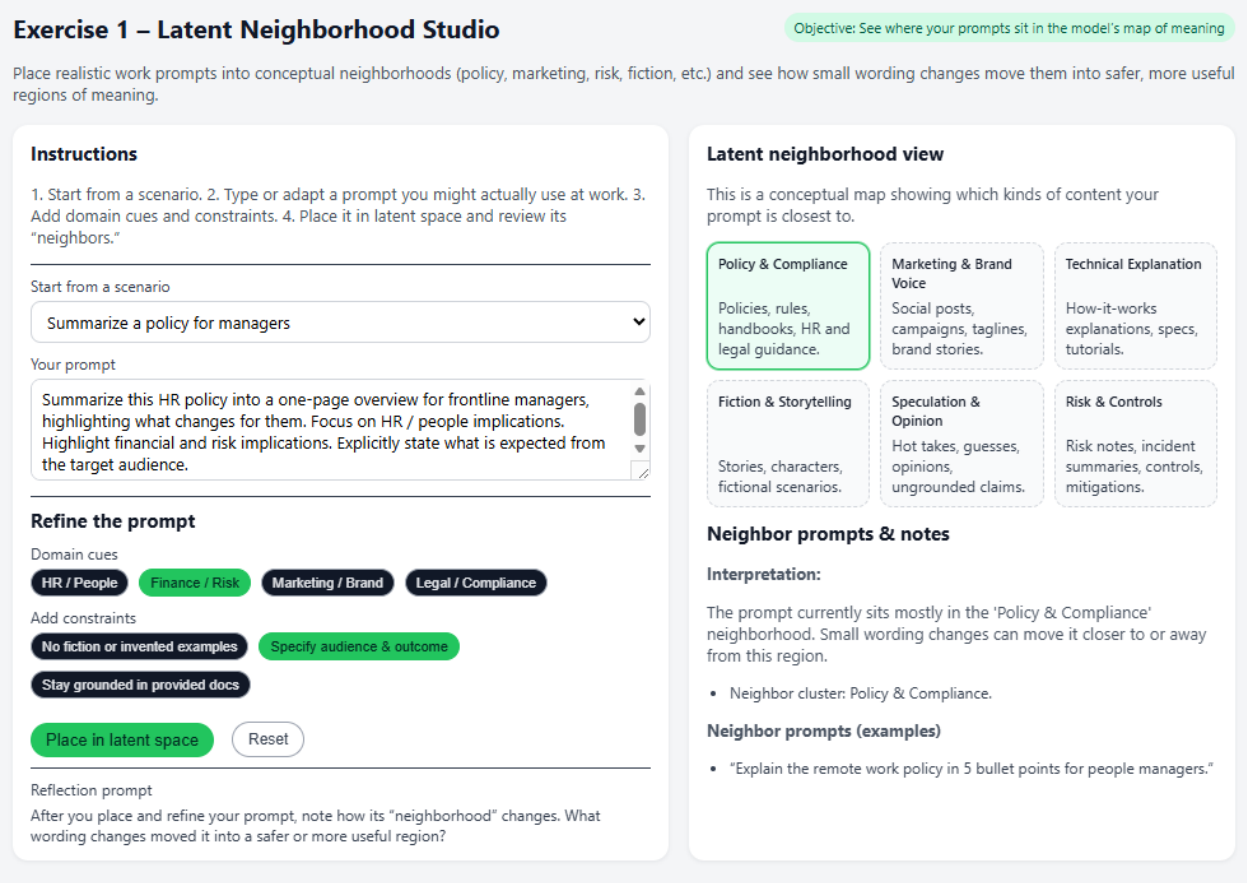


---

3. Exercise 1 – Latent Neighborhood Studio

Theme: Where does your prompt "live" in the model's map of meaning?

In the app:
Exercise 1 – Latent Neighborhood Studio



What you'll practice

You'll practice designing and refining prompts by **seeing which semantic neighborhood they fall into**:

- Policy & compliance
- Marketing & brand voice
- Technical explanation
- Fiction & storytelling
- Speculation & opinion
- Risk & controls

You move from:

- "Latent space is an abstract map of meaning"
  to

- “I understand where my real prompts land, and I can nudge them into a safer or more useful region.”

How to use it

1. **Choose a scenario**
    - Use the **Start from a scenario** dropdown:
        - Summarize a policy for managers
        - Draft a short marketing blurb
        - Write a vendor risk note
        - Write a short story opening
    - The tool pre-fills a sample prompt you can edit.
2. **Edit the prompt**
    - In **Your prompt**, adapt the text to something **you would actually ask an AI** in your job.
3. **Add domain cues and constraints**
    - Use the **Domain cues** chips (HR / Finance / Marketing / Legal).
    - Use the **Add constraints** chips (e.g., “No fiction or invented examples”, “Specify audience & outcome”).
    - These chips append extra guidance directly into your prompt so the model has clearer intent.
4. **Place it in latent space**
    - Click **Place in latent space**.
    - The **Latent neighborhood view** highlights 1–2 clusters where your prompt “lives.”
    - The **Neighbor prompts & notes** panel shows example neighbor prompts and a short explanation.
5. **Experiment and refine**
    - Change a few words (e.g., clarify the audience, add “grounded in provided docs”).

    - Click **Place in latent space** again and see if your neighborhood changes.

6. **Reflect**
    - Use the **Reflection prompt** to note:
        - What wording changes moved your prompt closer to Policy & Compliance vs Marketing & Brand vs Risk & Controls?
        - Is that where you actually want it?

How to get value from it

- Try prompts you already use at work and see if they land in **unexpected neighborhoods** (e.g., too speculative, too fictional).
- Use this exercise to answer:

"If the model thinks this is a marketing / fiction / speculation task, what could go wrong?"

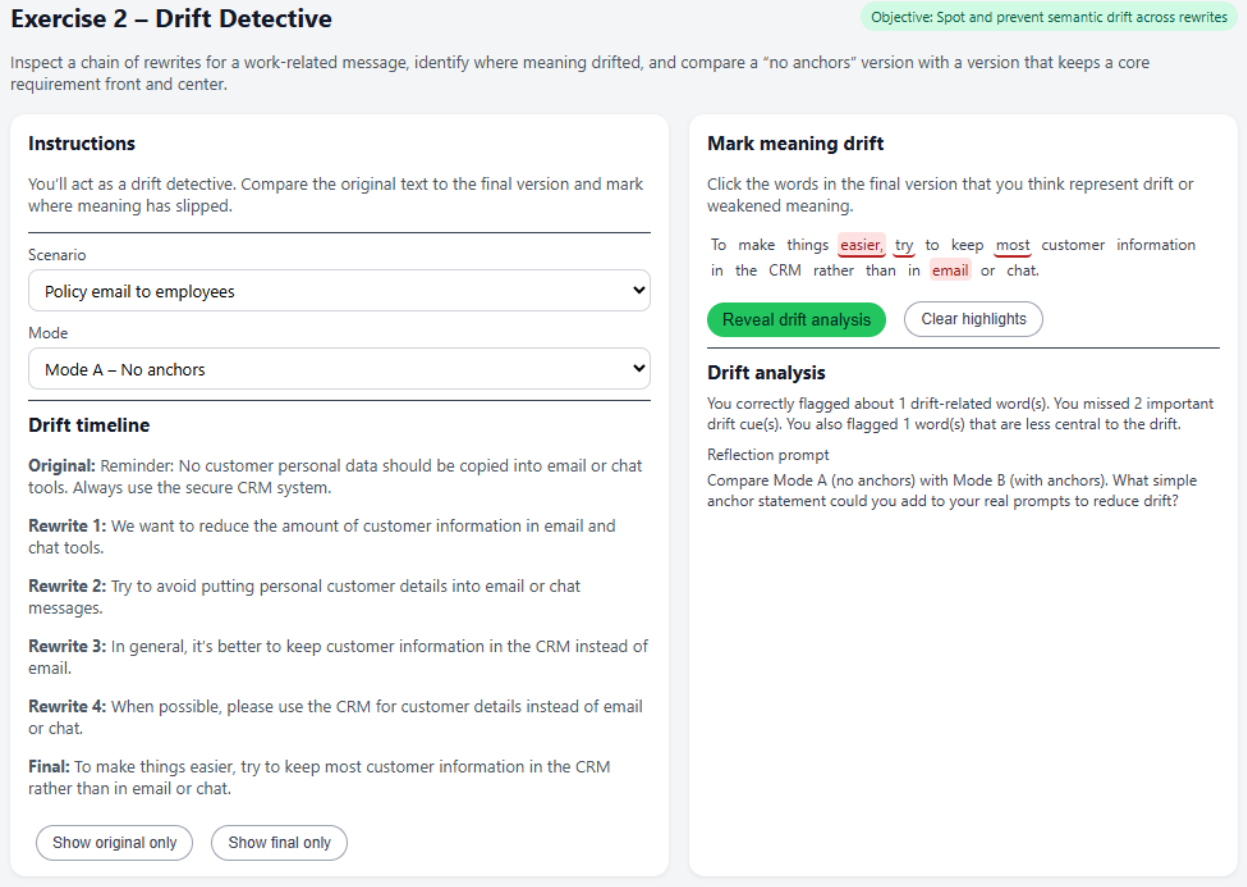
- Turn your insights into **prompt templates** for your team—each template intentionally anchored in the right neighborhood.

---

4. Exercise 2 – Drift Detective

Theme: How does meaning slowly slide away across rewrites?

In the app:
Exercise 2 – Drift Detective



What you'll practice

You'll practice spotting **semantic drift**—where a message becomes softer, less precise, or more permissive as it's rewritten.

You move from:

- "Drift is a theoretical risk"
  to
- "I can recognize where wording shifts turn hard rules into vague suggestions."

How to use it

1. **Pick a scenario**
    - Use the **Scenario** dropdown:
        - Policy email to employees
        - Client update memo
        - Incident report summary

2. **Choose a mode**
    - **Mode A – No anchors:** A rewrite chain where meaning drifts.
    - **Mode B – With anchors:** A rewrite chain where a short "core rule" is repeated to stabilize meaning.
3. **Review the drift timeline**
    - The **Drift timeline** card shows:
        - Original statement
        - Four intermediate rewrites
        - Final version
    - Read from original → final and notice subtle changes in strength (e.g., "must" → "generally" → "ideally").
4. **Mark drift in the final version**
    - In **Mark meaning drift**, the final text is broken into clickable words.
    - Click the words you think **weaken or change the original requirement** (e.g., "typically", "may", "exceptions").
5. **Reveal the drift analysis**
    - Click **Reveal drift analysis**:
        - You'll see how many drift-related words you caught.
        - The model highlights true "drift cues" in the text.
        - You get coaching on where strict rules turned into soft suggestions.
6. **Compare modes**
    - Switch between **Mode A – No anchors** and **Mode B – With anchors**.
    - See how a simple repeated phrase like

"[Core rule: No personal data in email or chat.]"
keeps meaning tight.

7. **Reflect**
    - Use the **Reflection prompt** to answer:

“What anchor phrase could I add to my real-world prompts and policies to stop this kind of drift?”

How to get value from it

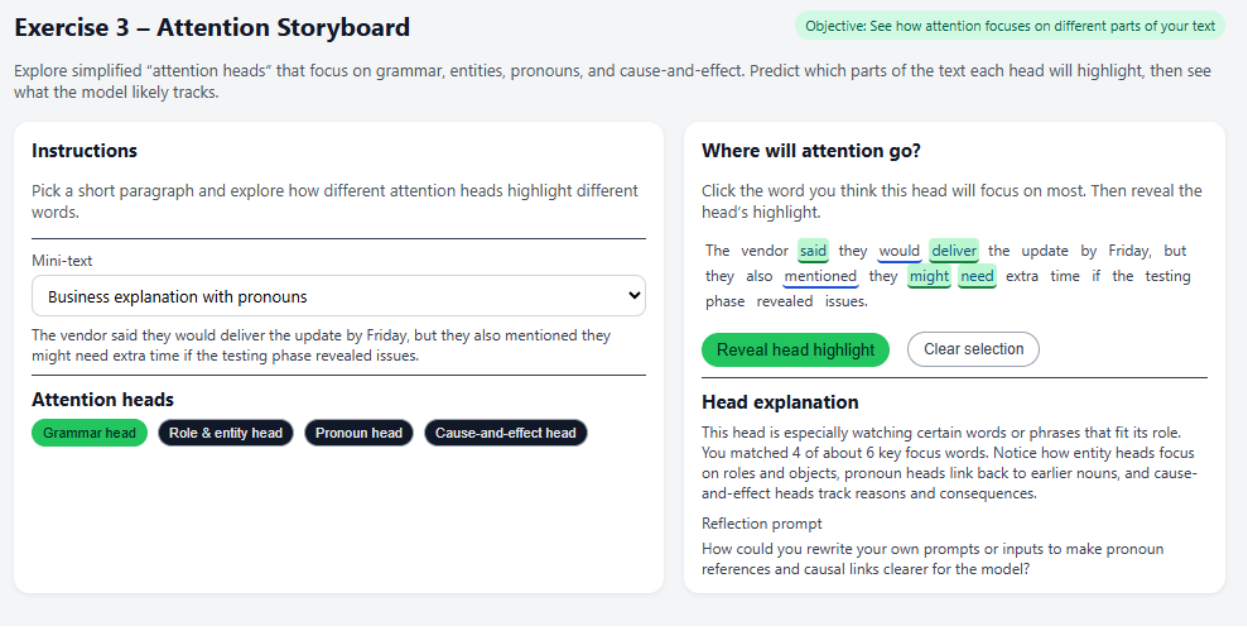
- This exercise trains your **risk radar** for subtle language softening in:
    - Policies
    - Compliance updates
    - Safety instructions
- Take your favorite anchor phrase from here and start **embedding it into your own prompts** for high-stakes tasks.

---

5. Exercise 3 – Attention Storyboard

Theme: What is the model actually paying attention to in your text?

In the app:
Exercise 3 – Attention Storyboard



What you'll practice

You'll practice thinking like an **attention head**:

- One head tracks grammar and verbs.
- One tracks entities (people, systems, dates).
- One tracks pronouns ("they", "it").
- One tracks cause-and-effect.

You move from:

- "Attention is a diagram in the book"
  to
- "I can predict which words the model needs to focus on—and rewrite to make that easier."

How to use it

1. **Choose a mini-text**
    - Use the **Mini-text** dropdown:
        - Short story-style paragraph
        - Policy explanation with conditions
        - Business explanation with pronouns
    - The text appears in **Mini-text** and as clickable tokens in **Where will attention go?**
2. **Pick an attention head**
    - Click a head chip:
        - Grammar head
        - Role & entity head
        - Pronoun head
        - Cause-and-effect head
    - This tells the app which kind of pattern you want to explore.

3. **Predict focus words**
    - In **Where will attention go?** click the words you think this head will focus on:
        - For pronoun head: pronouns and their references.
        - For cause-and-effect: words indicating reasons and consequences.
        - For entities: key roles, systems, dates.
4. **Reveal the head highlight**
    - Click **Reveal head highlight**.
    - The tool:
        - Highlights the “true” focus words for that head.
        - Marks where your guesses match.
        - Explains why those words matter.
5. **Reset and try another head**
    - Use **Clear selection** and switch to a different head on the same text.
    - See how each head views the text differently.
6. **Reflect**
    - In the **Head explanation** and **Reflection prompt** areas, note:
        - Where references or causes are ambiguous.
        - How a rewrite could make pronoun links or responsibilities clearer.

How to get value from it

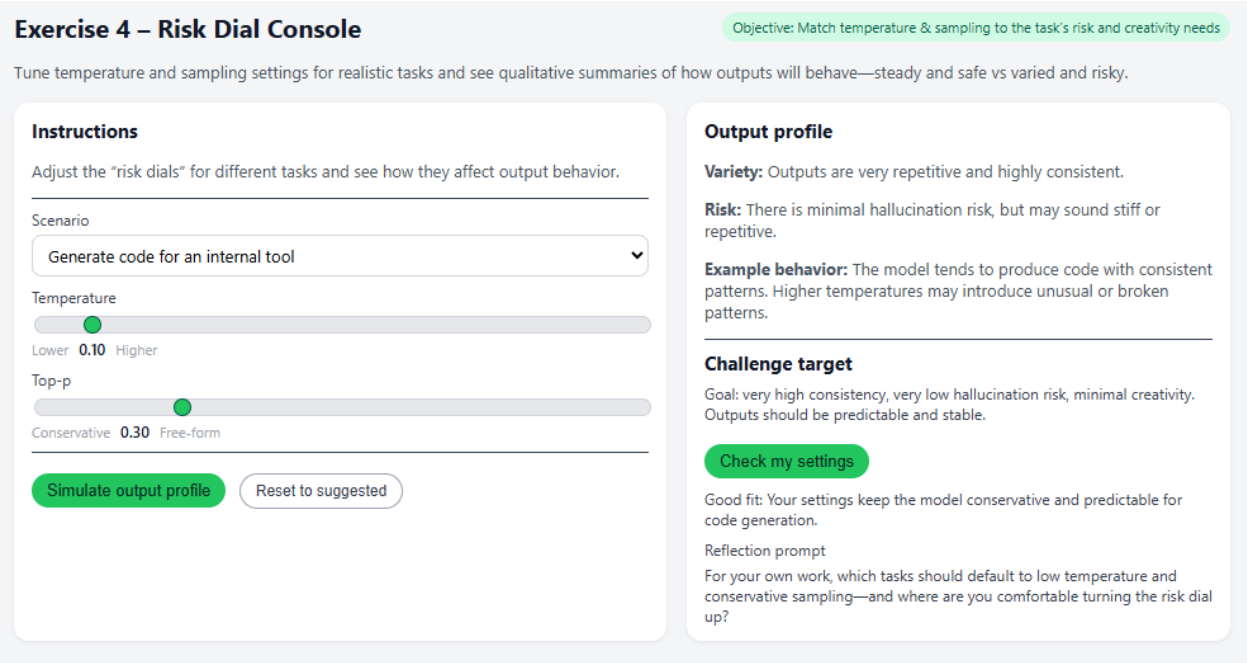
- You’ll see why vague phrases like “they” or “it” can confuse the model—and readers.
- Use what you learn to **rewrite your prompts and inputs**:
    - Name entities clearly (“the vendor”, “the security team”).
    - Spell out cause and effect.
    - Avoid unclear pronouns in critical instructions.

---

6. Exercise 4 – Risk Dial Console

Theme: Can you set the creativity vs reliability dials for the task?

In the app:
Exercise 4 – Risk Dial Console



What you'll practice

You'll practice tuning **temperature** and **top-p sampling** for different scenarios:

- Code generation
- Brainstorming
- Policy explanation
- Creative analogies

You move from:

- "Temperature is the model's randomness"
  to
- "I know which settings I'd choose for my real tasks—and why."

How to use it

1. **Select a scenario**
    - Use the **Scenario** dropdown:
        - Generate code for an internal tool
        - Brainstorm campaign ideas
        - Draft a customer-facing policy explanation
        - Write creative analogies for explaining embeddings
2. **Adjust the dials**
    - **Temperature slider:** 0.0 (very stable) to 1.5 (very exploratory).
    - **Top-p slider:** 0.1 (conservative) to 1.0 (free-form).
    - Labels beneath the sliders show the numeric values.
3. **Simulate output behavior**
    - Click **Simulate output profile**.
    - **Output profile** shows:
        - Variety: how repetitive vs varied the outputs will be.
        - Risk: how likely you are to see hallucinations, off-brand tone, or nonsense.
        - A short example of behavior in this scenario.
4. **Read the challenge target**
    - The **Challenge target** describes your goal for that scenario (e.g., “very high consistency, very low hallucination risk”).
5. **Check your settings**
    - Click **Check my settings**.
    - The tool tells you if your dials:
        - Match the target
        - Are too conservative
        - Are too adventurous

- It suggests better ranges for that task.

6. **Reset to recommended values**
    - Use **Reset to suggested** to see a recommended starting point for each scenario.
7. **Reflect**
    - In the **Reflection prompt**, answer:

"For my own work, which tasks should default to low temperature and conservative sampling, and where am I comfortable turning the dial up?"

How to get value from it

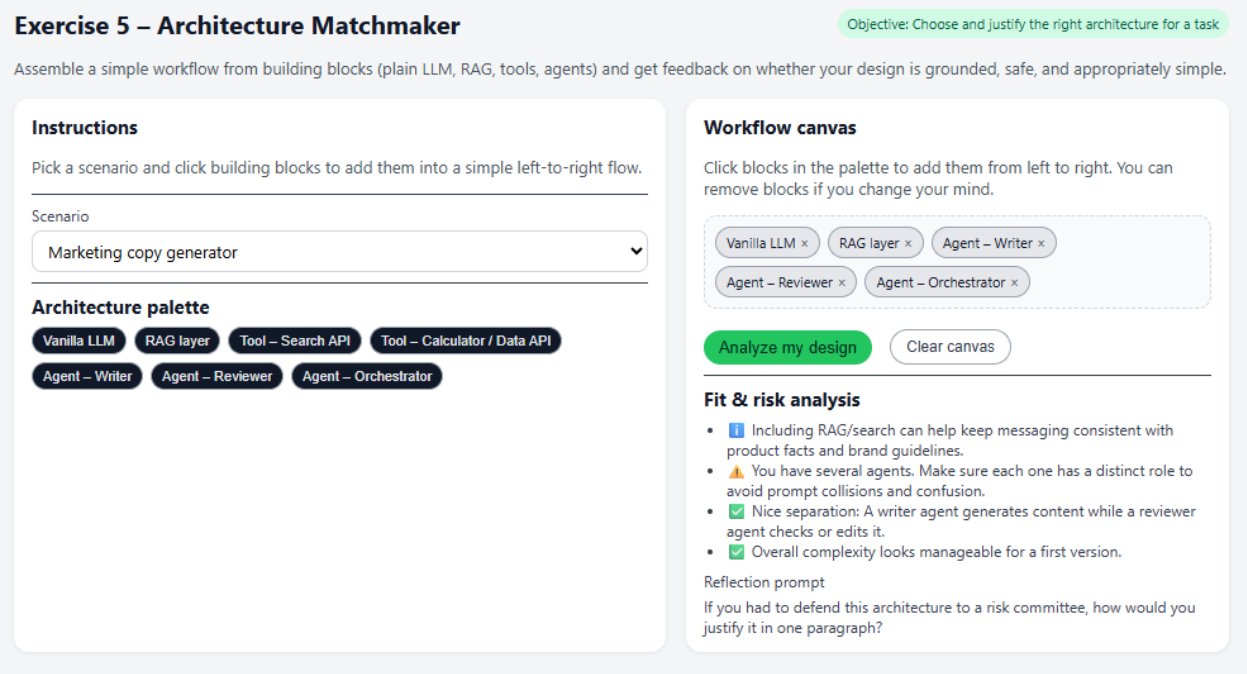
- For **high-risk tasks** (policy, compliance, external statements), you'll see why low temperature and conservative top-p are safer defaults.
- For **creative tasks** (brainstorming, analogies), you'll learn how to safely "open the faucet" for more variety without losing all control.
- Use this to create a simple internal guide:
    - "For task X, we recommend T ≈ 0.2–0.4, Top-p ≈ 0.3–0.6."

---

7. Exercise 5 – Architecture Matchmaker

Theme: Which architecture would you trust for this job?

In the app:
Exercise 5 – Architecture Matchmaker



What you'll practice

You'll design a **simple AI architecture** for each scenario by combining:

- Vanilla LLM
- RAG layer
- Tools (Search, Calculator/Data)
- Agents (Writer, Reviewer, Orchestrator)

You move from:

- "RAG, tools, and agents are labels"
  to
- "I can choose and justify a simple, safe design for a realistic use case."

How to use it

1. **Choose a scenario**
    - Use the **Scenario** dropdown:
        - Internal policy chatbot (needs grounding)
        - Marketing copy generator
        - Compliance Q&A assistant

        - Vendor risk review workflow

2. **Add building blocks**
    - In **Architecture palette**, click chips like:
        - Vanilla LLM
        - RAG layer
        - Tool – Search API
        - Tool – Calculator / Data API
        - Agent – Writer
        - Agent – Reviewer
        - Agent – Orchestrator
    - Each click adds that block to the **Workflow canvas** from left to right.
3. **Edit the canvas**
    - On the canvas, each block appears as a small "pill" with a **×** button.
    - Click **×** to remove a block if you change your mind.
4. **Analyze your design**
    - Click **Analyze my design**.
    - The **Fit & risk analysis** shows:
        - Whether you included **RAG** where grounding is critical (e.g., policy/compliance)
        - Whether you're relying on **plain LLM** in risky scenarios
        - Whether you have **too many agents** (over-complexity) or too few checks
        - Whether tools like **Search or Calculator** are missing where they'd help
5. **Clear and try again**
    - Use **Clear canvas** to start over with a different architecture for the same scenario or a new one.

6. **Reflect**
    - In the **Reflection prompt**, answer:

"If I had to defend this architecture to a risk committee, how would I justify it in one paragraph?"

How to get value from it

- You'll practice thinking like a **solution designer**, not a model builder:
    - When is **plain LLM** enough?
    - When is **RAG non-negotiable**?
    - When do you truly need multiple agents vs a simple setup?
- Use your favorite designs as starting points when you talk with IT, data, or vendor teams about **how** AI should be deployed—not just **if** it should.

---

8. Putting it all together

As you work through the Chapter 1 Exercise Lab:

1. **Collect your notes**
    - Save your reflections from each exercise:
        - How you changed prompts
        - Where you saw drift
        - What clarity helped attention
        - Which temperature/Top-p defaults feel right
        - Which architectures you trust
2. **Look for patterns**
    - How does **latent neighborhood** relate to:
        - Drift risk?
        - Attention patterns?

        - Architecture choices (e.g., when RAG is essential)?
    - Where are you personally most comfortable turning the **creativity dial up—and where not?**

3. **Turn insights into practice**
    - Write 3–5 simple rules for your own use of AI, for example:
        - "I will always specify domain, audience, and outcome in prompts for high-stakes tasks."
        - "For policy and compliance questions, I will prefer grounded designs (RAG + reviewer) over plain LLM."
        - "If wording turns 'must' into 'usually' or 'ideally' in safety or compliance areas, that's a red flag."
        - "For code and calculations, I keep the risk dial low; for idea generation, I'm okay with higher creativity."

Done well, these exercises turn Chapter 1 from "theory I've read" into **habits I can use in my daily work**—how I write prompts, read AI answers, and evaluate AI systems.