

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.



Chapter 1 Exercise Lab

Foundational & theoretical mastery

Exercise 1 – Latent Neighborhood Studio
Where your prompt “lives”

Exercise 2 – Drift Detective
Find meaning drift

Exercise 3 – Attention Storyboard
What the model “looks at”

Exercise 4 – Risk Dial Console
Temperature & sampling

Exercise 5 – Architecture Matchmaker
RAG, tools, agents

These exercises are for knowledge workers.
Use them to design, critique, and govern AI
behavior—not to write code.

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

Exercise 1 – Latent Neighborhood Studio

Objective: See where your prompts sit in the model's map of meaning

Place realistic work prompts into conceptual neighborhoods (policy, marketing, risk, fiction, etc.) and see how small wording changes move them into safer, more useful regions of meaning.

Instructions

1. Start from a scenario. 2. Type or adapt a prompt you might actually use at work. 3. Add domain cues and constraints. 4. Place it in latent space and review its "neighbors."

Start from a scenario

Summarize a policy for managers

Your prompt

Summarize this HR policy into a one-page overview for frontline managers, highlighting what changes for them. Focus on HR / people implications. Highlight financial and risk implications. Explicitly state what is expected from the target audience.

Refine the prompt

Domain cues

HR / People Finance / Risk Marketing / Brand Legal / Compliance

Add constraints

No fiction or invented examples Specify audience & outcome
Stay grounded in provided docs

Place in latent space Reset

Reflection prompt

After you place and refine your prompt, note how its "neighborhood" changes. What wording changes moved it into a safer or more useful region?

Latent neighborhood view

This is a conceptual map showing which kinds of content your prompt is closest to.

Policy & Compliance	Marketing & Brand Voice	Technical Explanation
Policies, rules, handbooks, HR and legal guidance.	Social posts, campaigns, taglines, brand stories.	How-it-works explanations, specs, tutorials.

Fiction & Storytelling	Speculation & Opinion	Risk & Controls
Stories, characters, fictional scenarios.	Hot takes, guesses, opinions, ungrounded claims.	Risk notes, incident summaries, controls, mitigations.

Neighbor prompts & notes

Interpretation:

The prompt currently sits mostly in the 'Policy & Compliance' neighborhood. Small wording changes can move it closer to or away from this region.

- Neighbor cluster: Policy & Compliance.

Neighbor prompts (examples)

- "Explain the remote work policy in 5 bullet points for people managers."

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

Exercise 2 – Drift Detective

Objective: Spot and prevent semantic drift across rewrites

Inspect a chain of rewrites for a work-related message, identify where meaning drifted, and compare a "no anchors" version with a version that keeps a core requirement front and center.

Instructions

You'll act as a drift detective. Compare the original text to the final version and mark where meaning has slipped.

Scenario

Policy email to employees

Mode

Mode A – No anchors

Drift timeline

Original: Reminder: No customer personal data should be copied into email or chat tools. Always use the secure CRM system.

Rewrite 1: We want to reduce the amount of customer information in email and chat tools.

Rewrite 2: Try to avoid putting personal customer details into email or chat messages.

Rewrite 3: In general, it's better to keep customer information in the CRM instead of email.

Rewrite 4: When possible, please use the CRM for customer details instead of email or chat.

Final: To make things easier, try to keep most customer information in the CRM rather than in email or chat.

Show original only

Show final only

Mark meaning drift

Click the words in the final version that you think represent drift or weakened meaning.

To make things easier, try to keep most customer information in the CRM rather than in email or chat.

Reveal drift analysis

Clear highlights

Drift analysis

You correctly flagged about 1 drift-related word(s). You missed 2 important drift cue(s). You also flagged 1 word(s) that are less central to the drift.

Reflection prompt

Compare Mode A (no anchors) with Mode B (with anchors). What simple anchor statement could you add to your real prompts to reduce drift?

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

Exercise 3 – Attention Storyboard

Objective: See how attention focuses on different parts of your text

Explore simplified “attention heads” that focus on grammar, entities, pronouns, and cause-and-effect. Predict which parts of the text each head will highlight, then see what the model likely tracks.

Instructions

Pick a short paragraph and explore how different attention heads highlight different words.

Mini-text

Business explanation with pronouns

The vendor said they would deliver the update by Friday, but they also mentioned they might need extra time if the testing phase revealed issues.

Attention heads

Grammar head Role & entity head Pronoun head Cause-and-effect head

Where will attention go?

Click the word you think this head will focus on most. Then reveal the head's highlight.

The vendor said they would deliver the update by Friday, but they also mentioned they might need extra time if the testing phase revealed issues.

Reveal head highlight Clear selection

Head explanation

This head is especially watching certain words or phrases that fit its role. You matched 4 of about 6 key focus words. Notice how entity heads focus on roles and objects, pronoun heads link back to earlier nouns, and cause-and-effect heads track reasons and consequences.

Reflection prompt
How could you rewrite your own prompts or inputs to make pronoun references and causal links clearer for the model?

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

Exercise 4 – Risk Dial Console

Objective: Match temperature & sampling to the task's risk and creativity needs

Tune temperature and sampling settings for realistic tasks and see qualitative summaries of how outputs will behave—steady and safe vs varied and risky.

Instructions

Adjust the “risk dials” for different tasks and see how they affect output behavior.

Scenario: Generate code for an internal tool

Temperature: 0.10

Top-p: 0.30

Simulate output profile | Reset to suggested

Output profile

Variety: Outputs are very repetitive and highly consistent.

Risk: There is minimal hallucination risk, but may sound stiff or repetitive.

Example behavior: The model tends to produce code with consistent patterns. Higher temperatures may introduce unusual or broken patterns.

Challenge target

Goal: very high consistency, very low hallucination risk, minimal creativity. Outputs should be predictable and stable.

Check my settings

Good fit: Your settings keep the model conservative and predictable for code generation.

Reflection prompt

For your own work, which tasks should default to low temperature and conservative sampling—and where are you comfortable turning the risk dial up?

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 $T \approx 0.2\text{--}0.4$, Top-p $\approx 0.3\text{--}0.6$.”

7. Exercise 5 – Architecture Matchmaker

Theme: Which architecture would you trust for this job?

In the app:

Exercise 5 – Architecture Matchmaker

Exercise 5 – Architecture Matchmaker

Objective: Choose and justify the right architecture for a task

Assemble a simple workflow from building blocks (plain LLM, RAG, tools, agents) and get feedback on whether your design is grounded, safe, and appropriately simple.

Instructions

Pick a scenario and click building blocks to add them into a simple left-to-right flow.

Scenario

Marketing copy generator

Architecture palette

Vanilla LLMRAG layerTool – Search APITool – Calculator / Data APIAgent – WriterAgent – ReviewerAgent – Orchestrator

Workflow canvas

Click blocks in the palette to add them from left to right. You can remove blocks if you change your mind.

Vanilla LLM xRAG layer xAgent – Writer x
Agent – Reviewer xAgent – Orchestrator x

Analyze my designClear canvas

Fit & risk analysis

- ! Including RAG/search can help keep messaging consistent with product facts and brand guidelines.
- ! You have several agents. Make sure each one has a distinct role to avoid prompt collisions and confusion.
- ✓ Nice separation: A writer agent generates content while a reviewer agent checks or edits it.
- ✓ Overall complexity looks manageable for a first version.

Reflection prompt
If you had to defend this architecture to a risk committee, how would you justify it in one paragraph?

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.