Interactive Labs User Guide

This guide explains how to run, navigate, and learn from the Chapter 1 Interactive Labs application. The app replaces/augments static Chapter 1 graphics with hands-on HTML5 simulations for latent space, embeddings, semantic drift, attention, and sampling controls.

1. What this application is

The application is a single-page, offline HTML5 learning environment. It uses small “toy” 2D embeddings and simulated attention to make Chapter 1 concepts tangible. The goal is intuition and skill-building, not reproducing a full LLM.

- Runs locally in any modern browser (Chrome/Edge/Firefox).
- No installation, servers, or external libraries needed.
- Five labs correspond to specific Chapter 1 sections.

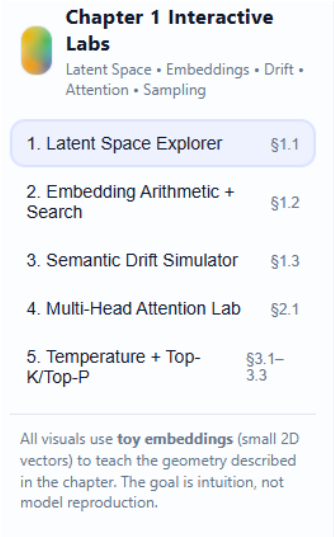
2. Getting started

2.1 Open the app

url

2.2 Navigation

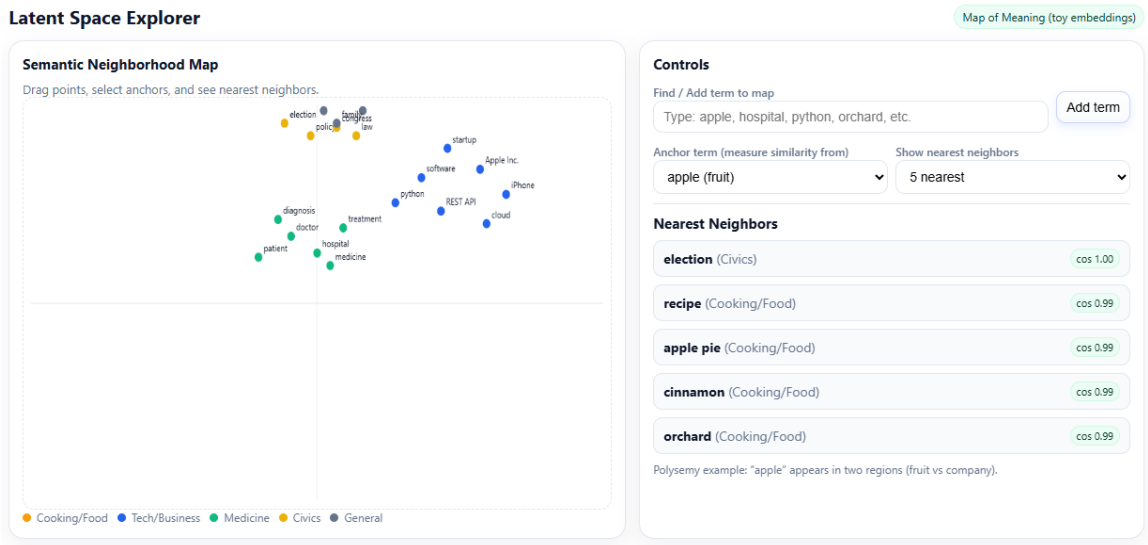
The left navigation menu lists the five labs. Click a lab title to switch views. On narrow screens the menu hides; scroll normally through each lab section.



3. Lab-by-lab walkthrough

Lab 1 — Latent Space Explorer **(Chapter §1.1)**

Purpose: Understand latent space as a “map of meaning,” where distance reflects semantic similarity and clusters represent related concepts.



How to use

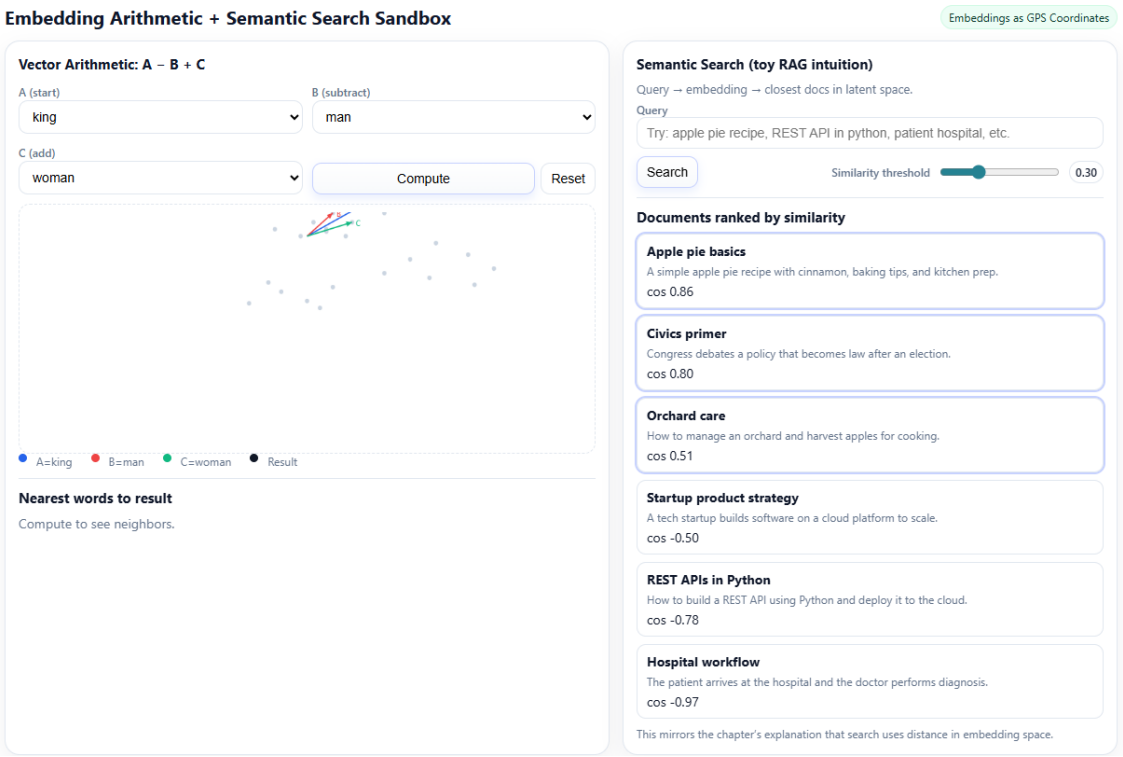
1. Look at the scatter map. Each dot is a word/phrase; color indicates a semantic group.
2. Click any dot to make it the Anchor (thick outline).
3. Use the ‘Show nearest neighbors’ dropdown to choose 3/5/8 neighbors.
4. Drag dots to see how moving a concept changes its neighborhood.
5. Type a term in ‘Find/Add term to map’ and click ‘Add term’ to focus on that concept.

How to interpret

- Nearby dots = similar meaning (high cosine similarity).
- Separate clusters = different semantic domains.
- Polysemy example: “apple (fruit)” and “Apple Inc.” sit in different regions—same surface word, different meanings.

Lab 2 — Embedding Arithmetic + Semantic Search **(Chapter §1.2)**

Purpose: Build intuition for embeddings as coordinates and for vector math producing meaningful analogies; understand semantic search as nearest-neighbor retrieval.



Vector Arithmetic panel

6. Pick A, B, and C using the dropdowns.
7. Click ‘Compute’ to calculate A − B + C.
8. Observe arrows on the plot: blue=A, red=B, green=C, black=result.
9. Read the ‘Nearest words to result’ list to see the best match.

Semantic Search panel

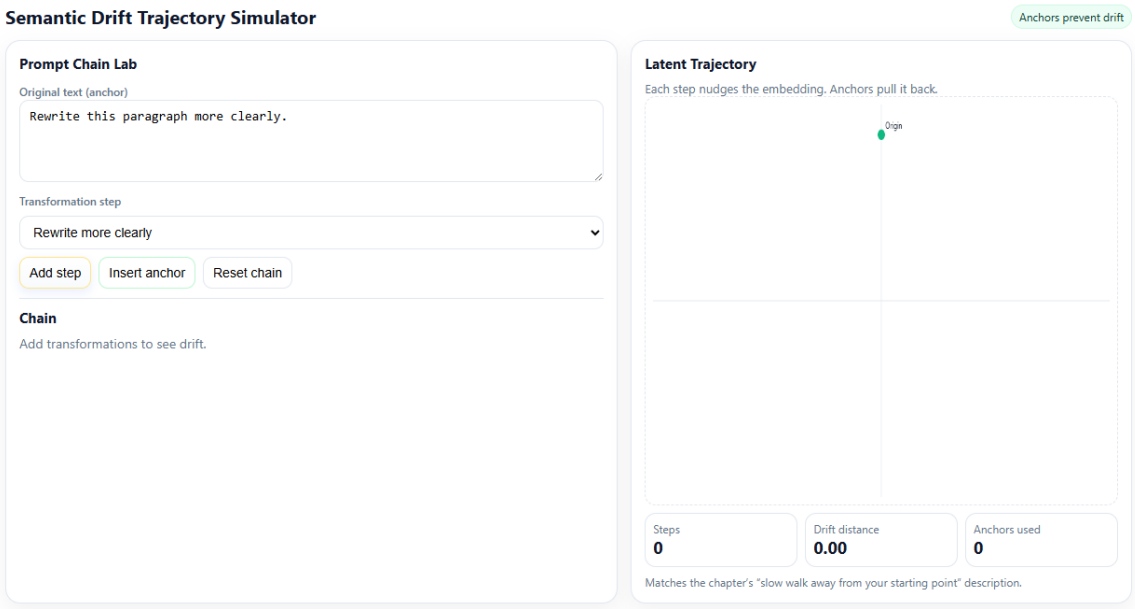
10. Enter a query in the Query box.
11. Adjust the similarity threshold slider (higher = stricter).
12. Click ‘Search’.
13. Documents are ranked by cosine similarity; those above threshold are highlighted.

How to interpret

- Vector arithmetic approximates relational meaning (e.g., king − man + woman ≈ queen).
- Semantic search retrieves content closer to the query in embedding space.
- Changing the threshold demonstrates precision/recall tradeoffs.

Lab 3 — Semantic Drift Trajectory Simulator (Chapter §1.3)

Purpose: See drift as accumulated meaning shift across iterative transformations and learn how anchors pull outputs back toward intent.



How to use

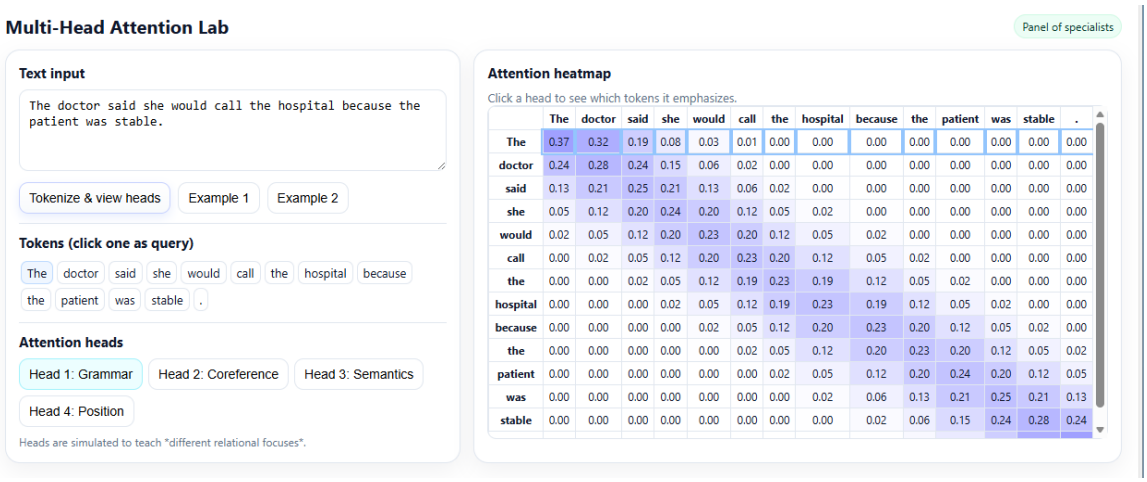
14. Enter the original text in ‘Original text (anchor)’.
15. Choose a transformation step from the dropdown (clarify, formal, concise, etc.).
16. Click ‘Add step’ repeatedly to simulate multi-turn prompting.
17. Watch the trajectory move away from the Origin point.
18. Click ‘Insert anchor’ to restate constraints; trajectory snaps back.
19. Use ‘Reset chain’ to start over with the same origin.

How to interpret

- Each blue node is a transformation step; the line shows cumulative drift.
- Orange ‘Anchor’ steps reduce distance from Origin.
- The KPI cards summarize steps, drift distance, and anchors used.

Lab 4 — Multi-Head Attention Lab **(Chapter §2.1)**

Purpose: Experience multi-head attention as parallel ‘specialists’ focusing on different relationships in the same text.



	The	doctor	said	she	would	call	the	hospital	because	the	patient	was	stable	.
The	0.37	0.32	0.19	0.08	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
doctor	0.24	0.28	0.24	0.15	0.06	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
said	0.13	0.21	0.25	0.21	0.13	0.06	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
she	0.05	0.12	0.20	0.24	0.20	0.12	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00
would	0.02	0.05	0.12	0.20	0.23	0.20	0.12	0.05	0.02	0.00	0.00	0.00	0.00	0.00
call	0.00	0.02	0.05	0.12	0.20	0.23	0.20	0.12	0.05	0.02	0.00	0.00	0.00	0.00
the	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00	0.00	0.00
hospital	0.00	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00	0.00
because	0.00	0.00	0.00	0.00	0.02	0.05	0.12	0.20	0.23	0.20	0.12	0.05	0.02	0.00
the	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.12	0.20	0.23	0.20	0.12	0.05	0.02
patient	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.12	0.20	0.24	0.20	0.12	0.05
was	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.13	0.21	0.25	0.21	0.13
stable	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.15	0.24	0.28	0.24

How to use

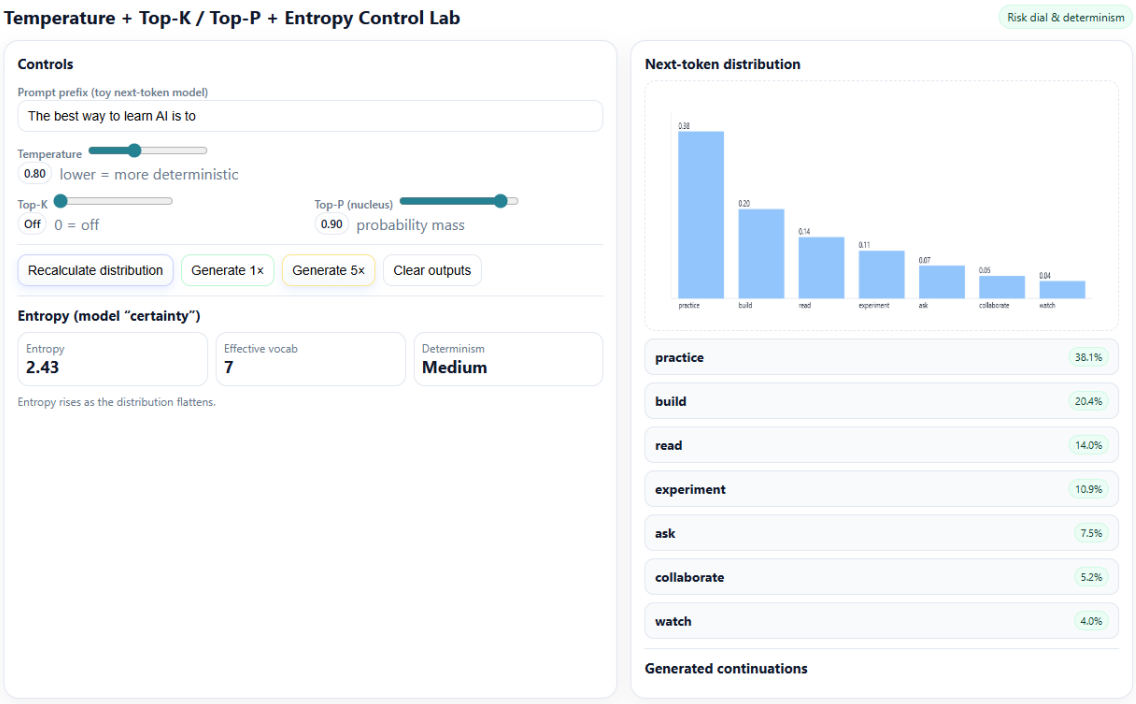
20. Type or paste a sentence into the text box.
21. Click ‘Tokenize & view heads’. Tokens appear below.
22. Click a token to highlight its attention row.
23. Click different Head buttons to switch specialist views.
24. Use Example 1/2 to load pre-made sentences.

How to interpret

- The heatmap shows how strongly each token attends to others (numbers are probabilities).
- Different heads emphasize different patterns: grammar proximity, pronoun coreference, semantic links, or positional structure.
- Comparing heads illustrates why multiple heads capture richer context than one.

Lab 5 — Temperature + Top-K/Top-P + Entropy Lab (Chapter §3.1–§3.3)

Purpose: Understand decoding controls as sliders that reshape next-token probability distributions, affecting creativity, consistency, and certainty (entropy).



How to use

25. Set Temperature with the slider. Lower = safer/more deterministic; higher = more random/creative.
26. Optionally set Top-K (number of candidates) and/or Top-P (probability mass).
27. Click 'Recalculate distribution' if you want a manual refresh (sliders auto-refresh too).
28. Inspect the bar chart and ranked list of candidate tokens.
29. Click 'Generate 1×' or 'Generate 5×' to sample outputs repeatedly and observe variance.
30. Use 'Clear outputs' to reset the generation list.

How to interpret

- A peaked distribution (low entropy) yields similar outputs each run.
- A flat distribution (high entropy) yields more varied outputs.
- Top-K and Top-P prune the tail of low-probability tokens.
- The entropy meter is a visible proxy for confidence/uncertainty.

4. Recommended learning activities

- Latent Space: Pick an anchor term and explain why its top 5 neighbors make sense; then drag it into a different cluster and observe changes.
- Embeddings: Try 3 analogies; note which succeed and which fail, and why.
- Drift: Build a 5-step chain; measure drift; insert anchors until drift falls below 0.5.
- Attention: Identify which head best resolves pronouns in your sentence.
- Sampling: For one prompt, compare Temperature 0.3 vs 1.2 with Top-P on/off; write a rule of thumb for each setting.

5. Troubleshooting

- Nothing shows when opening index.html: try Chrome/Edge, or use the local server method.
- Text looks too small: use browser zoom (Ctrl + / Ctrl −).
- Want to embed in a site: upload the folder to GitHub Pages or any static host; keep index.html at root.

6. Concept mapping back to Chapter 1

- §1.1 Latent Space → Lab 1
- §1.2 Embeddings & Semantic Search → Lab 2
- §1.3 Semantic Drift → Lab 3
- §2.1 Multi-Head Attention → Lab 4
- §3.1–3.3 Temperature, Sampling, Entropy → Lab 5